KU LEUVEN
FACULTEIT LETTEREN
ONDERZOEKSEENHEID TAALKUNDE

Modelling interactive alignment

A multimodal and temporal account

Proefschrift voorgedragen tot het behalen van de graad van
Doctor in de Taalkunde

door

Bert Oben

Promotor: prof. dr. Geert Brône
Co-promotor: prof. dr. Kurt Feyaerts

2015

Dankwoord

“Iets met humor was het toch he?”. Dat is wat de meeste mensen denken dat ik de afgelopen jaren heb onderzocht. De tijd is nu eindelijk aangebroken dat die mensen zelf onomstotelijk, empirisch kunnen vaststellen dat dat niet klopt. Ik heb me de voorbije jaren bezig gehouden met het bestuderen van kopiegedrag tijdens face-to-face gesprekken. Dat werk was soms grappig, soms lachwekkend en tegen het einde niet gelachen, maar met humor heeft het niets te maken. Laat me dus de mensen bedanken die me de afgelopen jaren verhinderd hebben een doctoraat rond humor te schrijven.

Kurt Feyaerts. Jij bent de reden waarom mensen denken dat dit werk over humor zal gaan. Zeven jaar geleden nam je me aan om een vak over humor en creativiteit op de rails te helpen zetten. Dat onderwijsproject hield onder andere het maken en beheren van een corpus van spontane gesprekken in. Daar is mijn interesse voor corpusgebaseerd onderzoek naar fenomenen in spontane conversaties begonnen. Samen met Geert hield je me in dienst, zelfs met de laatste resten van de CHIL-clubkas, waardoor ik nog steeds in beeld was toen er in 2012 een zak FWO-zilverlingen als hemels manna uit de lucht kwam vallen. Voor dat vertrouwen en dat doorzettingsvermogen ben ik jullie heel dankbaar. Kurt, je erg diverse wetenschappelijke interesse en je lekker koppig-kritische ingesteldheid hebben dit doctoraat sterk beïnvloed. Bedankt om telkens weer je indrukwekkend aantal bezigheden aan de kant te schuiven en ongebreideld tijd te maken om diep door te bomen over methodologische of theoretische kwesties. Bedankt ook om tijdens de laatste maanden rust en vertrouwen uit te stralen en naast de immer zinvolle en kritische opmerkingen ook geregeld “yes”, “alright” of “hell yeah!” als opmerking toe te voegen.

Geert Brône. Je speltechnische capaciteiten in het snooker staan in schril contrast tot je academische capaciteiten. Je bent in beiden even

gedreven, maar academisch gezien speel je het spel op het hoogste niveau mee, en met veel succes. Ik heb het geluk om mee te kunnen surfen op je onpeilbare drive van enthousiasme, belezenheid en ondernemerschap. Je steile carrière is moeilijk te evenaren, maar wel een duidelijk en groot voorbeeld. Het is indrukwekkend hoe je me door je brede kennis en kritische geest (en niet aflatende speelse innuendo) telkens weer scherp houdt. Daarnaast is het een ongehoorde luxe dat een promotor iemand is die tot je vriendenkring is gaan behoren. Kurt, Geert, ik ben ontzettend blij en dankbaar dat onze academische maar ook onze persoonlijke paden elkaar hebben gekruist.

Lieve ouders. De pagina's na dit dankwoord gaan misschien grotendeels aan jullie voorbij, dus lees deze pagina maar des te vaker. Bedankt om me met alle kansen en in de meest comfortabel mogelijke omstandigheden aan de start van mijn academisch parcours te brengen. Zonder jullie zou ik niet over de nodige rust, gezond verstand en doorzettingsvermogen beschikken om dit verhaal tot een goed einde te brengen.

Lieve schoonouders. In de eerste plaats, bedankt om Annick te maken. Dat was ontzettend vriendelijk. Maar verder ook bedankt om me het leven makkelijker te maken in drukke tijden. Dankzij jullie heb ik minder tijd aan klussen in huis en verversen van luiers moeten spenderen. Zoals jullie nu kunnen vaststellen, heb ik die voor mij vrijgemaakte tijd zinvol besteed.

De mannen van De Raaskalderij. Ik heb de afgelopen tijd veel meer academisch dan ambachtelijk gekalde raas geproduceerd. Het is tijd om dat evenwicht te herstellen. Bedankt om me te blijven meeslepen in jullie heerlijke en niet-aflatende onnozelheid.

Rik. Jij wist vaak beter aan welk hoofdstuk en welk onderdeel ik aan het werken was dan ikzelf. Op risico van door hamburgervet gevoede levercirrose verteerde je elk detail van mijn vorderingen en iedere klaagzang zonder verpinken. Bedankt daarvoor.

Nog niet genoemde familie en vrienden. Het feit dat jullie generisch worden aangesproken, maakt jullie niet minder belangrijk. Net het uitgebreide netwerk van kleine duwtjes in de rug en ondersteunende

handjes maken dat ik in erg comfortabele omstandigheden mijn nerdschap heb kunnen bedrijven om dit doctoraat te schrijven. Zus, bedankt ook voor het last-minute en long-distance naleeswerk.

Collega's. Geen doctoraat zonder koffie. Geen koffie zonder klets. Bedankt om de stoffigheid van onze betonnen bunker weg te blazen met jullie enthousiasme, dwaze spelletjes en oprechte peilingen naar de vorderingen van onderhavig werk. Liesbeth en Elisabeth: bedankt voor het dagelijkse klankbord op kantoor; Jelena, Steven en Paul: ondanks een naamsverandering ben ik blij dat de sfeer in onze kleine groep nog steeds op en top *chil* is.

Nina. Je beschikt nog niet over voldoende cognitieve vaardigheden om het zelf al te lezen, maar ik schrijf het toch. Je bent de copernicaanse omwenteling van mijn persoonlijke universum. Je hebt mezelf uit het centrum van mijn wereld weggeslingerd en hebt daar autoritair, edoch lichtjes kwijlend, zelf plaats genomen. Geheel en al terecht. Bedankt om me met je dwingende schattigheid er geregeld aan te herinneren dat dit doctoraat er helemaal niet toe doet.

Annick. Bovenal, Annick. Het is geniaal hoe wij met zijn tweetjes aligneren op alle multimodale niveaus die ik me kan inbeelden en wensen. De brompot die wat in zichzelf gekeerd zit te mijmeren, het warhoofd dat alweer twee afspraken tegelijkertijd heeft gepland, de afwezige omdat hij op congres of voor drie maanden in Tilburg zit: bedankt om met al die Berten om te gaan. En om ze nog graag te zien ook. Maar nog veel meer bedankt om er iedere dag weer voor te zorgen dat mijn hartslag wat hoger gaat als ik je na een dag werken weer zie. Het klinkt romantisch-melig, maar ik meen het wel: alleen mijn liefde voor jou is nog groter dan mijn dankbaarheid. Zonder jou geen doctoraat. Zonder jou geen echte Bert. Zonder jou helemaal niks!

Contents

Introduction

My dad is a true chameleon. He cannot change colours, but he is remarkably good at changing dialects. If I bump into my dad when he is talking to someone on the telephone, I can reliably guess who he is talking to just by listening to his dialect imitation. He lives in a rural area where every tiny village still has its own and very specific dialect. Whenever he is on the phone, he copies the dialect of his conversational partner. This copying occurs at different levels: he uses words, constructions and phonetic realisations that are not indigenous to his own dialect, but to that of his interlocutor. This type of copying behaviour is what we will label *alignment* and will be the phenomenon under scrutiny in this dissertation.

To give some concrete examples of the phenomenon of alignment, and how this can occur at different levels, consider the examples below. The playful banter in (1) is an exchange of letters between the British statesman Winston Churchill and the Irish writer George Bernard Shaw. It illustrates how conversational partners can gear their utterances to one another at multiple linguistic levels: Churchill copies some of Shaw's words, but crucial to the joke, he also copies Shaw's grammatical conditional construction with an if-clause. Whether this tiny slice of communication really ever took place, is a matter of debate, but the communicative pattern is clear: Churchill very consciously copies his conversational partner's lexical and grammatical means. It is certainly no coincidence that Churchill uses those exact words and that specific conditional if-construction: the linguistic alignment helps him to attain the conversational goal of trumping his friend Shaw.

(1)

"I am enclosing two tickets to the first night of my new play
Bring a friend ... if you have one."
— *George Bernard Shaw (in a letter to Winston Churchill)*

"I cannot possibly attend the first night.
I will attend the second … if there is one."
— *Churchill's response note*

Apart from being an enjoyable, witty piece of human interaction, the example in (1) is very well suited to explain which aspect of alignment we will and will not study in this dissertation. First, the example consists of written interaction. In the present study we will only consider spoken, face-to-face conversations. Second, because spoken discourse requires instantaneous linguistic actions and reactions, speakers have less time to cleverly retort than the protagonists in (1). This sparseness of reaction time does not mean, however, that people do not draw on their partner's linguistic input. To the contrary, recycling linguistic elements from a conversational partner is a very efficient mechanism to cope with the constraining cognitive load during face-to-face interaction. As a result, the alignment we measure will not only encompass semantically rich, humorous and witty pairs of utterances, but also very automatic, often unconscious types of alignment such as using the same intensifiers, diminutives or passive constructions as your interactional partner. Third, the present study will focus on more than the linguistic level, and incorporate behavioural alignment as well. More specifically, we add gaze, gesture and intonation to our analyses.

Not taking such a multimodal perspective on the topic of alignment would rule out important aspects of what actually happens when speakers engage in spontaneous talk. This is clear from the example in (2)[1]: if we only take into account the transcript, there is no alignment to be observed between the two protagonists. However, by looking at the corresponding video, we clearly see Baldric copying general Melchett's intonation and cheek pinching gesture.

(2)

"Are you looking forward to the big push?"
— General Melchett (to Baldric)

"No sir, I am absolutely terrified."
— Baldric's response

VIDEO: http://www.youtube.com/watch?v=IDQ1ljlnSjU&t=1m24s

[1] Example taken from the BBC comedy Blackadder, season 4, episode 1.

As was the case in (1), the example in (2) illustrates a very deliberate and conscious kind of alignment that is aimed at humorously trumping the conversational partner. In the present study we will also, and in fact mainly, study more automatic types of multimodal alignment such as alignment of gaze direction, hand shapes or vocal tone frequencies. What we will not study is types of alignment that are not formally observable. Speakers might, for example, align in terms of using irony or in terms of adopting an overly formal and archaic register. This type of alignment is not directly formally measurable from the transcriptions or the audio-visual data and will not be considered in this dissertation.

The observation that alignment occurs at many different levels is far from new. A large body of research has provided convincing evidence of that: speakers copy each other's lexical choice (Brennan & Clark 1996, Garrod & Anderson 1987), prosodic features (Giles & Powesland 1975, Lewandowski 2012, Szczepek Reed 2010), syntax (Branigan et al. 2007, Gries 2005) or use of indirect language (Roche et al. 2010). Also beyond the verbal level interlocutors appear to align. Chartrand and Bargh (1999) showed that if subjects see their conversational partner rub their nose or shake their foot, they will unconsciously imitate that behaviour. This alignment has also been demonstrated for headshakes, nods, laughter and eyebrow raising (Louwerse et al. 2012), posture (Shockley et al. 2003), and even for heart rates (Konvalinka et al. 2011).

What is generally lacking in research on the topic, is how alignment at different levels is interconnected, and when during conversations interlocutors align more (or less). These two issues make up the core of the present study. We do not want to add to the existing literature another multimodal layer at which people appear to align. Rather, we start from the basic observation that interlocutors are not aligned all the time at any possible level. We want to show how alignment at one level is linked to alignment at another. For example, we will study whether lexical and gestural alignment co-occur: if interlocutors use the same word to refer to an object, will they also align gesturally in referring to that object? This will be the *multimodal* dimension of this study. Second, we also want to study how alignment unfolds dynamically over time: is there more alignment as people interact longer? Or do we see local peaks of alignment rather than a

gradual increase? Or is alignment that automatic that we observe no *temporal* effect at all? To allow both a multimodal and temporal take on the phenomenon of alignment, it will be crucial to find good measures of identifying what counts as alignment. Are two gestures with a different palm orientation but the same hand shape aligned? Is there still alignment if there is more than eight minutes between an identical prime (what speaker 1 does) and target (what speaker 2 does)? Or how do we measure gaze alignment: by counting the number of *times* interlocutors look at the same thing, or by counting the number of *seconds* they do so? These issues all come under the *methodological* dimension of this dissertation.

In Chapter 1 we will first present a state of the art in which we show different approaches to the same phenomenon of alignment. Because speech, gesture and gaze are the levels under scrutiny in this dissertation, we review studies on interactive alignment in these modes of representation. After demonstrating *that* people align at different levels, we also discuss *why* they might do so. Based on this literature overview, we position this study and formulate our research questions. In Chapter 2 we first briefly review existing multimodal corpora, then argue why we need a more specific corpus than the ones available, to finally describe the Insight Interaction Corpus that will be the basis for all of the analyses in this dissertation. In Chapter 3 we take a multimodal perspective towards alignment and present two case studies that show which factors are good predictors of lexical and gestural alignment. The perspective is multimodal because we look into how one multimodal level (i.c. eye gaze) affects alignment at other levels (i.c. lexical and gestural alignment), and because we check whether the same factors predict alignment at different multimodal levels. In Chapter 4 the perspective is not multimodal but temporal. A first case study demonstrates how interlocutors synchronise their gaze behaviour. In a second study we illustrate the temporal dynamics of alignment, i.e. whether it occurs in local peaks and whether it increases over interaction time. Chapter 5 is an integration of a multimodal and temporal perspective because we study how the temporal dynamics of alignment at one level are linked to that at another level. In a correlation analysis we demonstrate which multimodal levels appear to globally correlate in terms of alignment rates. In a cluster analysis we show for

which multimodal levels we measure co-occurring peaks of alignment rates. In the conclusion we first sum up the results and link them back to the existing studies and theories described in earlier chapters, and finally present some suggestions for future research.

Chapter 1

State of the art
& Research questions

People align their behaviour at many different levels. In the example in the introductory chapter we saw how Churchill playfully imitates Shaw at the lexical and syntactic level. Studying how interlocutors use linguistic alignment to achieve humour (as in Brône & Oben 2013) requires a different approach and yields a different theoretical result than studying how people align their behavioural mannerisms such as nose scratching or feet rubbing (as in Chartrand & Bargh 1999). Because alignment is such a multi-level phenomenon, it has been studied from an equally wide range of scientific disciplines. In this chapter we do not want to provide an *exhaustive* overview of studies on alignment, but we do try and offer a *structured* overview of which disciplines are involved in studying which topics. After a schematic state of the art, we will position the present study in the current research field and formulate our research questions.

1.1 Terminological issues

One of the reasons why providing a structured overview of research on alignment is not straightforward, is the terminological fuzziness surrounding the phenomenon. Researchers have coined many different names for the observation that speakers copy each other's behaviour: shadowing (Goldinger 1998, Lewandowski 2012), resonance (Brône & Zima, 2014, Du Bois 2014), entrainment (Garrod & Anderson 1987), accommodation (Giles et al. 1991), structural priming (Bock & Griffin 2000; Howes, Healey & Purver 2010), conceptual pacts (Brennan & Clark 1996), parallelism (Sakita 2006; Tannen 1987, 1989), mimicry (Kimbara 2006), convergence (Michelas & Nguyen 2012), the 'Chameleon effect' (Chartrand & Bargh 1999), adaptation (Brennan & Hanna 2009, Mol et al. 2012), coordination (Fusaroli et al. 2012; Richardson, Dale & Kirkham 2007; Tolston et al. 2014), etc. The problem is that these different names also imply different approaches and theoretical presuppositions, and more importantly, they do not refer to the exact same phenomenon.

In this study we use *alignment* as a cover term to refer to any formal repetition of behaviour across speakers[1]. Alignment in our definition thus always involves a behaviour by a first speaker (which we call *prime*) followed by that same behaviour by a second speaker (which we call *target*). We use the term *alignment* regardless of the exact temporal relation between prime and target, regardless of the intentionality of the observed behaviour, and regardless of the type of behaviour (gaze, gesture, speech, posture, etc.). Although the term is inspired by the influential work by Pickering and Garrod (2004, 2006) on *interactive alignment*, it is important to make explicit that we do not want to slavishly endorse the theoretical claims that come with the interactive alignment theory. The discussion sections in chapters 3, 4 and 5 will make clear where we con- and diverge from the interactive alignment theory.

Central to the present study is the multimodal perspective on alignment. We have just pointed out that this multimodal alignment goes by many different names. In addition, the term *multimodal alignment* itself can also refer to different phenomena[2]:

i. indicating simultaneity or other temporal relations between events on different multimodal layers;
ii. referring to copying behaviour across speakers on different multimodal layers.

Multimodal alignment in (i) typically refers to the temporal relation or co-occurrence of gesture and speech (Campana et al. 2005; Hadar 2013; Kopp, Bergmann & Wachsmuth 2008), speech and gaze (Cummins 2012, Hadelich & Crocker 2006), or gesture and gaze (Gullberg & Kita 2009, Oben & Brône,

[1] As already pointed out in the introductory chapter: we do not consider pragmatic types of alignment (e.g. irony as in Roche, Dale & Gaucci 2010) or implicit types of alignment like (e.g. syntactic reinterpretations as in Zima 2013), but only instances of alignment that are directly, formally observable in the transcriptions or audio-visual data.

[2] Some researchers also use *alignment* in terms of affiliation. For example, for Stivers (2008: 32) addressees *align* if they "acknowledge the information provided in the telling and support the progress of the telling" by their conversational partner.

forthc.) in the production of multimodal utterances. *Alignment* here expresses a mere temporal relation between events in different semiotic channels, regardless of any copying behaviour. This copying behaviour is crucial in (ii), where *multimodal alignment* indicates that the copying occurs at different multimodal levels.

1.2 Conceptual issues

Not only terminological issues complicate a clear communication on the topic of alignment. Before we move to a literature overview, we also want to disentangle some conceptual issues like the relationship between the notions of *alignment*, *synchronisation* and *coordination*, and the interplay between *intentionality* and *semantics*.

1.2.1 Alignment and synchronisation during coordination

According to Clark (1996) conversation is a form of *joint action*. Just like a married couple trying to assemble an IKEA wardrobe, conversational partners negotiating a car sale need to coordinate their actions (i.e. their language use) to achieve their goal. If the couple in the IKEA example does not coordinate its actions, i.e. if they perform actions like screwing, nailing or hammering regardless of what the other is doing, the result will be a marital quarrel instead of a practical and reasonably-priced wardrobe. The same goes for the car sale negotiation: if the speakers involved utter sentences independently from one another, there will be no deal. Even more, without any form of linguistic coordination we can hardly say there is a conversation at all. These examples make clear that it not only matters what we say (or do), it also matters a great deal when and how we say (or do) it. This issue touches upon some key concepts we want to clarify before moving on: how does *coordination* relate to *alignment* (*what we do*) and *synchronisation* (*when we do it*)?

The overview in Fig. 1 starts from the assumption that spoken interaction (like any other act of coordination) can either involve alignment or not. Either both partners do the same thing, or they do not. Applied to the car sale example: either both seller and buyer refer to the price in terms of “quid”, or the one uses “quid” and the other “pounds”. Conversations in which none of the words by one speaker are used by the other (and vice

versa) are unlikely, but nonetheless possible. We consider actions or utterances that are coordinated but not aligned, as *complementary*. Crucial here is that the actions or utterances occur in a specific temporal order: the car seller cannot start the conversation with "it was nice doing business with you, sir", nor can he end with "welcome to our showroom". To come to a sensible, efficient and teleologically oriented interaction, both partners need to temporally organise their individual contributions to the conversation well. If they do, and in doing so not formally align to their partner, we label their behaviour as *behavioural attunement* (see Fig. 1). If they do not, thus in cases where there is neither alignment nor a temporal attunement of individual contributions, the question arises whether we can still consider this behaviour to be interaction (hence the question mark in Fig. 1).

Apart from *complementary behaviour*, conversational partners can also formally align to their partner while speaking. In terms of the car sale example, both buyer and seller might use the connective "suppose that" to start hypothetical scenarios. This alignment can occur without any temporal relation between the occurrences of "suppose that" (which would be *behaviour matching* in Fig. 1). If there is a temporal dependency between the aligned behaviour of the two speakers, we consider that to be *behavioural synchrony*. This temporal relation can be manifold. There can be a relation of *convergence*: e.g. at the beginning of the car sale the buyer and seller use different connectives, but at the end they systematically use "suppose that". Another possibility is *parallelism* between the behaviour of the interactional partners: if the buyer starts talking louder, the seller will do the same thing. As a consequence the loudness of the conversation will fluctuate, but both speakers are equally loud throughout. A final type of temporal relation between aligned behaviour occurs when there is a systematic time *lag* between the aligned events: if the buyer yawns, the seller will yawn five seconds later. To sum up on the three types of interpersonal synchrony, and applying it to the case of yawning during spoken interaction: if speakers start yawning more towards the end of a conversation, they *converge*; if they yawn at the same time there is *parallelism*, and if their yawning consistently occurs with five seconds in between, there is a *time lag* (see Fig. 1).

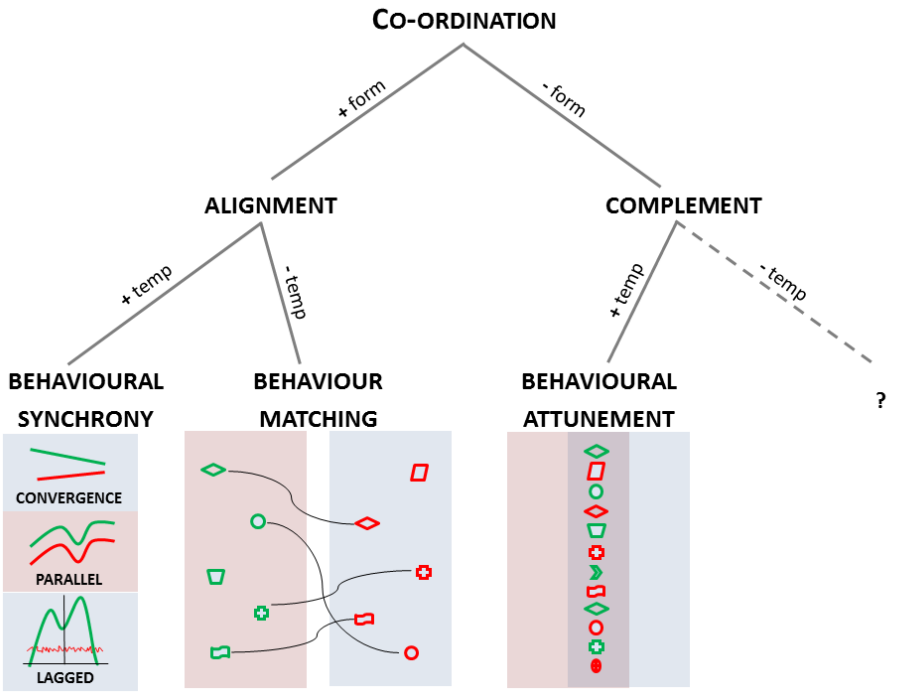

Fig. 1: A schematisation of the relation between coordination, alignment and synchronisation

1.2.2 Intentionality and semantics

Most of the research on alignment focusses on the automatic and mechanistic nature of the phenomenon (Chartrand & Bargh 1999; Lakin, Chartrand & Arkin 2008; Louwerse et al. 2012; Pickering & Garrod 2006). Some researchers (Branigan et al. 2000, Brennan & Clark 1996) do not commit themselves to a clear positioning in attributing conscious awareness to alignment (or not). Judging alignment of nose scratching to be unintentional and the humorous banter between Shaw and Churchill[3] as intentional is quite straightforward. Other examples are less self-evident to place in either category. Take for example referential gestures. If interlocutors align in using a drawing gesture[4] to refer to a door, how can

[3] See example (1) in the introductory chapter.

[4] See more on gesture types in section 2.3.5 of the next chapter. A drawing gesture means the fingers are used to trace the outline of the represented object.

we claim whether or not this alignment is intentional or not. In cases where there is an explicit meaning negotiation (i.e. interlocutors are talking about the drawing gesture itself), we can assume the gestural alignment is intentional. However, and this is the case most of the time, the lack of explicit meaning negotiation is a poor indicator of unintentionality. If there is no meaning negotiation, we cannot safely conclude the two gestures were performed unintentionally. This means there is a considerable *grey zone* between intentional and unintentional behaviour (see shaded part in Fig. 2).

A second dimension included in Fig. 2 is that of *semantics*. The speech and gesture behaviour under scrutiny in this dissertation either carries independent, semantic content (typically nouns, adjectives, verbs, depictive gestures, etc.), or not (typically backchannels, prepositions, pronouns, deictic gestures, etc.). The distinction between the two is more clear cut (hence no shaded parts and no overlap between the two in Fig. 2), compared to the intentional-unintentional dichotomy.

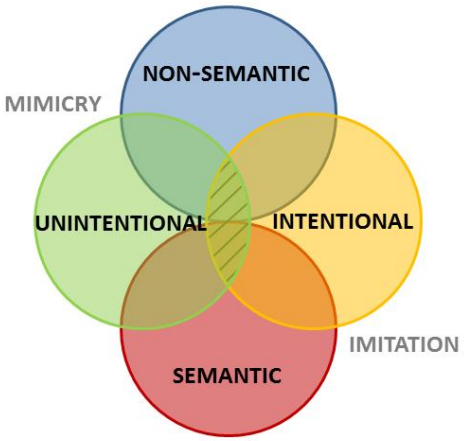

Fig. 2: A schematisation of the interaction between intentionality and semantics

In combining the notions of intentionality and semantics, we see two types of research on alignment (although other combinations are of course possible, yet they are studied less). On the one hand, researchers that focus more on unintentional and non-semantic behaviour, and on the other hand research on intentional, semantic types of behaviour. We will label the

former as *mimicry* and the latter as *imitation*. In what follows, we will use the concepts defined in this section on conceptual issues to give a state of the art of the current literature on the phenomenon of alignment.

1.3 A structured overview of research on alignment

1.3.1 Alignment: a pervasive phenomenon at multiple levels

Research on alignment is far from new. For example, the observation that in language learning (both from a phylogenetic and ontogenetic perspective) alignment is crucial, has captured researchers' attention over different centuries (Dominey 2004, James 1878, Piaget 1932). From the onset of scientific reports on the topic, alignment has been shown to occur at multiple levels, including non-verbal behaviour (Darwin, 1890/2009:36). In this section we will summarize contemporary research on alignment at different levels.

LEXICAL ALIGNMENT

Whenever speakers lexically label the world surrounding them, they have at their disposal a vast repository of possibilities. For example, when referring to their wives, husbands have myriads of ways to do so: "wife", "spouse", "love", "Catherine", "baby", "honey", "she", "you know who", etc. In making a referential choice, parameters such as informativeness, availability, cognitive load or perceptual salience play a role (Vogels 2014). The presence of two women named Catherine, disfavours the use of "Catherine" for the husband to refer to his wife. Similarly, when referring to a red car that is standing next to a van, a truck and a bicycle, speakers would tend not to use a label like "vehicle", because it is not informative enough for referential purposes (distinguishing one vehicle from the other), or "red car", which is too informative in the given situation. Rather, they should opt for the informatively most efficient and concise "car". However, *ahistorical* factors such as conciseness and efficiency are not the only factors determining referential choice. Brennan & Clark (1996) demonstrated that *historical*, contextual factors play a role too. Clark and colleagues adopt a strong interactional approach that takes partner-specific conceptualisations or *shared conceptualisations* as a driving force in dialogue. In other words, reference is argued to be designed to a large

extent with regard to the past interaction with co-participants. The starting point for this line of reasoning is the observation that out of context, there is still a high degree of variability in the linguistic construal of events or objects, despite the range of ahistorical principles (cited in Brennan & Clark 1996: 1483). Although conditions are identical for all the speakers, in a picture naming task some speakers will use "car", and others will use "red convertible" to refer to the same photograph. However, speakers that are engaged in interaction tend to use the same word as their interactional partners, even if that word is overinformative or underspecified, a point further proven by Goudbeek & Krahmer (2012).

The experiments in Brennan & Clark (1996) demonstrate that conceptualisation in interaction is subject to a process of *interactive grounding*: specific sets of partners reach a temporary agreement about a given lexical construal. These *conceptual pacts* are not directly transferable to other new addressees. The upshot is a strong interactional account of conceptualisation: speakers and addressees jointly set up conceptual pacts or shared conceptualisations for the purpose of the ongoing interaction, which result in local lexical routines. That this is a dynamic process had already been pointed out by Garrod & Anderson (1987). Based on a data set of subjects playing a maze game, they found conversational partners grow routines in indicating their position in the maze (descriptions using coordinates as in "I'm on C-4" vs path descriptions as in "See the bottom right, go two along and two up. That's where I am."). According to the authors, this *growing* of routines can be taken literally because they observed that this type of alignment was progressive, i.e. it increased as the interlocutors talked longer to each other.

Building on Garrod & Anderson (1987), Pickering & Garrod in their seminal work on *interactive alignment* (2004, 2006) provided further evidence of lexical alignment. Interestingly, they propose quite a different interpretation to the exact same observation than Brennan & Clark (1996) did. To Pickering & Garrod (2004, 2006) the process of alignment is automatic, i.e. it is almost entirely based on the input-output mechanism of priming. Basically, speakers use a word because they have just encountered it. Using a different word would be suboptimal in terms of cognitive effort: in conversation there is no room for constantly modelling the beliefs of

your conversational partner. For example, if a speaker uses "red convertible" to refer to a car in a photograph, the addressee activates the lexical items "red" and "convertible" during the comprehension of that utterance. Because of this rise in activation, addressees will be more likely to use "red convertible" themselves in subsequent language production. Perhaps it is needless to repeat, but it is crucial to keep this in mind, this priming operation runs independently from any other-modelling. According to Pickering & Garrod (2006: 221) the process of activation is "automatic and does not involve a conceptual pact between the interlocutors".

The difference between both interpretations of the same observation (i.e. two interlocutors using the same word) boils down to the difference between *grounding* and *priming*. The interactive alignment theory (Pickering & Garrod) assumes that cross-speaker alignment does not presuppose *shared* conceptualisation, i.e. priming enables interlocutors to efficiently align mental representations without having to tap into additional cognitive resources and without having to model each others' mental states. This priming mechanism is so fundamental that in many cases it would require more cognitive resources to override the basic tendency to align than to adhere to it (Costa, Pickering & Sorace 2008). To Brennan & Clark other-modelling is crucial in building *common ground*. Speakers design their utterances drawing on previous interaction(s) with their partner, either in the current conversation or even beyond. Alignment to them is not an individual cognitive mechanism, but rather a collaborative process not restricted to the individual mind. The tension between the two accounts of alignment is also linked to what in Fig. 2 has been marked as a grey zone between intentional and unintentional behaviour. For Pickering & Garrod, lexical alignment should clearly be placed in the domain of unintentional behaviour[5], whereas Brennan & Clark consider it at least at the intersection between intentional and unintentional.

[5] Although they do mention that intentional alignment is possible, for example in situations where experts consciously adapt their linguistic choices when talking to novices. They also report cases of intentional *non-alignment*, for example politicians consciously and consistently using either "terminate a pregnancy" or "murder a child" to refer to the same event of abortion.

Although Pickering & Garrod and Brennan & Clark disagree on some fundamental points, their studies can be placed on the same end of the semantic dimension in Fig. 2: both focus on semantic behaviour, i.e. they study how interlocutors use nouns, adjectives, verbs or more complex constructions to describe objects or events. What they also have in common is the type of data they use as a gateway into the phenomenon of alignment. Both use experimentally controlled conversations of speakers involved in a *collaborative task*, and even more specifically, a task within the director-matcher paradigm[6]. A different and more recent line of research diverges from this approach in (at least) two ways: they study corpus data of naturally occurring interactions and they are more interested in the non-semantic linguistic behaviour.

In recent years, researchers have paid attention to alignment of function words next to content words. According to Manson et al. (2013: 419), who also quote Ireland & Pennebakker (2010), function words are especially interesting because they are "inherently social". The comprehension of function words is more dependent on contextual information during speech, than on stable conventions and meanings. When talking about a "red convertible" a speaker can draw on the assumption his addressee knows what "red" and "convertible" mean. However, when talking about "the red convertible", speech partners must have established which car exactly they are talking about. Observing that both speakers use "red convertible" can be explained by the interlocutors' individual knowledge of the language, but observing alignment of "the red convertible", including the definite article, indicates the interlocutors share a frame of reference beyond their linguistic knowledge. Hence, compared to content words, function words are equally good, or maybe even better indicators of coordination processes among interlocutors.

[6] In this paradigm participants are attributed a role during a task. Typically the participants take turns in describing routes or (abstract geometrical) objects (*instruction givers)*, while their conversational partner (*instruction followers*) follows the route or identifies the shape from a series of possible alternatives.

What Manson et al. (2013) found was that speakers generally align their use of function words during a collaborative task[7]. However, not all function words reached significant alignment (e.g. conjunctions and quantifiers). Outside of the experimentally controlled lab setting, Danescu-Niculescu-Mizil et al. (2012) looked at oral arguments before the United States Supreme Court. In this setting of interaction between nine justices and the lawyers for both sides, they found that lawyers align more to justices (in terms of function words) than the other way around. Moreover, the researchers demonstrated a link between the alignment behaviour and the trial outcome: lawyers aligned more to justices that would end up voting against them (because they are more dependent on them, Danescu-Niculescu-Mizil and colleagues argue) than to justices voting in favour of them. Beňuš et al. (2014) add to these observations that, the other way around, justices do not align more or less to the lawyers they eventually vote for.

Alignment of function words has not only been demonstrated in face-to-face conversations, but also in written interaction. A corpus survey of interactions on Twitter (Danescu-Niculescu-Mizil, Gamon & Dumais 2011) revealed significant alignment for nearly all the function words under scrutiny. Interaction on Twitter is quite different from face-to-face conversation, and even other types of online messaging communication, because there is no real-time interaction (i.e. there is ample time to consciously plan and produce 'utterances') and there is a 140 character limit. In fact, Twitter was not designed to be a medium of conversation at all, which makes the observation of alignment all the more striking: alignment appears to be a sufficiently robust mechanism that can withstand the constraints imposed by the medium of Twitter. A different story is presented in Riordan et al. (2011). In an instant messaging corpus they only found significant alignment for two types of function words, i.c. pronouns and conjunctions. Whether the difference in alignment rates between the study on Twitter and that on instant messaging actually depends on the difference in medium, remains an issue to be resolved.

[7] All of the studies on function words reported in this section use the Linguistic Inquiry and Word Count (LWIC) tool or parameters (Pennebaker, Booth & Francis, 2007) to perform their analyses.

SYNTACTIC ALIGNMENT

Speakers not only re-use each other's words, also larger (and more abstract) chunks of speech appear to be aligned across speakers. Levelt & Kelter (1982) were one of the first to systematically study syntactic alignment. They found syntactic similarity in question-answer pairs: when people were asked questions like "In which car does Zed drive?" they would typically answer with "In a convertible". When the question was "Which car does Zed drive", an answer without the preposition "in" was most frequent. Weiner & Labov (1983) found that if speakers have a choice between a passive and an active sentence construction, chances are significantly higher they will produce the passive construction if there is another passive construction in the immediately preceding (i.c. a range of 5 clauses) discourse. More recently, in a large scale corpus study on the British National Corpus (BNC), Szmrecsanyi (2005) broadened the scope of syntactic patterns that are subject to alignment. Szmrecsanyi reports alignment of future tense markers ("I will drive my convertible" vs "I'm going to drive my convertible"), particle placement ("I pimped up my convertible" vs "I pimped my convertible up") and analytic versus syntactic comparatives ("My convertible is fancier than yours" vs "My convertible is more fancy than yours"). Moreover, he also demonstrated that the textual *distance* between prime and target is a relevant factor: the less words in between, the higher the chance that the two instances will be aligned. The results in Gries (2005) are parallel to those in Szmrecsanyi (2005). Gries uses the British component of the International Corpus of English (ICE-GB) and adds that also choices for either option in the dative alternation ("My dad gave me a convertible" vs "My dad gave a convertible to me") appears to be aligned. More importantly, Gries showed that self-repetition was stronger than other-repetition, and that priming effects are verb-specific. This means that speakers, for this syntactic feature of dative alternation, align more to themselves than to their conversational partners (although they still significantly align to their partners), and that some verbs seem to be more resistant or prone to alignment than others. A more abstract take on syntactic alignment was adopted by Dale & Spivey (2006). In a number of corpora of child-caretaker interaction they found alignment of word-class n-grams. The type of alignment they measure makes abstraction of any

specific syntactic construction. Combinations of words (i.c. 2-grams, 3-grams and 4-grams) pertaining to the same word-class seem to occur in each other's conversational neighbourhood. What is more, the directionality of this type of alignment is dependent on the proficiency level of the child. Younger and less proficient children copy their caregivers more (in terms of word-class n-grams) than the other way around, but if children grow older and more proficient this system tilts and caregivers become the *followers*, copying their *leading* child more often than vice-versa.

Apart from Levelt & Kelter (1982) all of the research on syntactic alignment reported so far is based on corpus linguistic analyses of spontaneous speech. However, syntactic alignment has also been intensively studied in experimental research. Interesting in this respect is a study by Reitter, Moore & Keller (2006) who compared syntactic alignment in both spontaneous speech (Switchboard corpus, Marcus et al. 1994) and experimentally generated data (Map Task corpus, Anderson et al. 1991). What Reitter and colleagues demonstrate, and this ties in with Gries (2005), is that self-repetition is present in both corpora, but other-repetition only in the task-oriented dialogue of the Map Task corpus. What was consistent across both corpora was the effect of the factor distance: the less distance there is between prime and target, the higher the chance they will be aligned. This distance effect played a role in both self-repetition and other-repetition.

A landmark publication for experimental research on syntactic alignment is Bock (1986). She provided proof that syntactic alignment, i.c. active-passive alternation and dative alternation, does not "depend on superficial relationships between successive sentences, but on more abstract structural similarities" (Bock 1986: 379). To be more precise, she uncovered that syntactic alignment occurs regardless of the lexical, conceptual or discourse content of the prime and target, regardless of the animacy of the agents in prime and target, and regardless of word order in prime and target. In a follow-up study Bock & Griffin (2000) focussed on the persistence of syntactic alignment. Their results indicate that syntactic alignment is not affected by the factor distance. In a span ranging from one to ten utterances, the authors found no significant decline in alignment. In other words, even though prime and target are ten utterances apart, they

still tend to be aligned. This is a remarkable result in that it contradicts the findings of Reitter, Moore & Keller (2006) and Szmrecsanyi (2005) who found the exact opposite.

Adding to the observations already outlined, a number of studies found an interesting interplay between lexical and syntactic alignment, i.e. the so-called *lexical boost effect*. Among others, Branigan, Pickering & Cleland (2000) and Cleland & Pickering (2003) found syntactic alignment was more likely to occur when prime and target contain the same (or conceptually related) lexical items. Note this is diametrically opposed to what Bock (1986) found. Branigan, Pickering & Cleland (2000) demonstrated that interlocutors in a director-matcher game tend to use syntactic structures[8] they had just encountered, but they significantly did more so if the verb in the constructions under scrutiny remained the same. Similarly, Cleland and Pickering (2003) show that alignment of noun phrase form (relative clauses like “The car that's red" vs pre-nominal clauses like "The red car”) occurred more often after a noun phrase using the exact same noun or a semantically related noun (“The truck that’s red”), in comparison to a semantically unrelated noun (“The room that’s red”). In contrast to an effect for semantically related nouns, the authors found no difference in alignment rates when prime and target nouns were phonologically related (“The car that’s red” and “The bar that’s red”) versus when they were unrelated (“The car that’s red” and “The room that’s red”).

More recently, Howes, Healey & Purver (2010) used corpus data of spontaneous speech to support the *lexical boost effect*. More importantly, what they demonstrate is that the lexical effect was so strong, they no longer measure an effect of syntactic alignment: “a match in syntactic structure across turns appears to be accounted for by the repetition of specific words” (pp. 2008). In line with Gries (2005) the authors also found that self-repetition is reliably stronger than other-repetition, but more crucial is the method they used to conclude that syntactic alignment did not occur more often than chance in their corpus. Howes and colleagues investigated prime-target pairs for the use of prepositional phrase (“Dad

[8] In this study by Branigan, Pickering & Cleland (2000) the use of double-object (“Dad gave me a convertible”) vs prepositional-object (“Dad gave a convertible to me”) structures was examined.

gave a convertible to me") versus double-object phrase ("Dad gave me a convertible") constructions. Suppose now that all the speakers in the corpus always use the construction with a preposition. In that case, a possible conclusion might be that speakers in the corpus are syntactically aligned all the time. To avoid this undesirable conclusion (i.e. to exclude there is no real choice between different constructions), the authors created *control dialogues*[9]. By randomizing the order of the utterances of each conversation, or by randomizing across speakers and thus combining transcriptions of speakers that never actually talked to each other, they created fake dialogs. Measuring syntactic alignment in these fake dialogs served as a control data set. Only if there is significantly more alignment in the real data set than in the fake one, it can be excluded that the alignment measured (in the real data) is due to chance. Apparently, based on their data set (i.e. the Diachronic Corpus of Present-Day Spoken English, DCPSE) there is no significant difference in syntactic alignment between random and real data. In a follow-up study Healey, Purver & Howes (2014: 5) generalize their findings from the transitive verb constructions to any syntactic construction, leaving us with quite a dissonant note in the *scene* on syntactic alignment: in contradiction to what most other researchers advocate, these authors claim that "people do not repeat their own or each other's syntactic structures more than would be expected by chance".

Most of the studies reported above can be placed in a psycholinguistic tradition. Within a functional/cognitive linguistic tradition the observation of structural cross-turn parallelism inspired researchers to the concept of *dialogic syntax* (Du Bois 2014, Sakita 2006). Within this theory, not the copying itself, but the effect it resorts (called *resonance*) is crucial. The focus is on *why* speakers reproduce parts of their partners' previous utterances and *how* this serves their communicative goals (Zima 2013). Opposed to 'traditional' syntax, dialogic syntax stretches over the boundaries of sentences or utterances and makes the parallelism itself a part of the grammar of spoken interaction. In a traditional syntax, grammatical rules define how smaller constituents can be combined to jointly make sense. In dialogic syntax, meaning arises from the

[9] For a more elaborate explanation of creating control dialogues, see section 3.2.2 in the third chapter.

juxtapositioning of parallel linguistic forms. Not only the internal structure of the individual utterances, but also the effect of the parallelism across utterances is what creates meaning. Going back to the letter conversation between Churchill and Shaw in the introductory chapter, only a grammar that takes into account the utterances of both gentlemen at the same time can fully describe the meaning of Churchill's retort. When only considering 'traditional' grammatical rules, a sentence such as "I will attend the second night if there is one" does not contain elements of humour, interpersonal stance or trumping. When considering the parallelism between Churchill's and Shaw's utterance, we can observe that a surplus in meaning (an effect of *resonance*) arises, which is crucial to the understanding of the interaction. Within the framework of dialogic syntax, it is claimed that speakers are very sensitive to and actively drawing on this effect of resonance during conversation. This is why structural cross-speaker parallelism can result in very local, ad-hoc constructions (as described in Brône & Zima 2014), and why it is such a frequent phenomenon.

GESTURAL ALIGNMENT

In section 1 of this chapter we pointed at some terminological issues regarding the term "alignment". Equally, the term "gesture" deserves some clarification. To some researchers it refers to any bodily movement, (including head movements, posture and facial expressions), to others it only applies to co-speech gestures performed with the arms and hands. In this study we will consistently use "gesture" in the latter sense. This is only a matter of terminology: we do not consider hand and arm movements as being more important or relevant than other bodily movement, we only use a separate name for it.

Research on gesture (with Kendon (1988, 2004) and McNeill (1992, 2005) as pioneers in the field) and by extension research on gestural alignment is a fairly young discipline. Kimbara (2006, 2008) performed conversation analysis (CA) research on the phenomenon and found people copying (parts of) each other's gesture during explicit grounding or meaning negotiation. One of the situations in which she observed gestural alignment was that of co-construction, i.e. situations where one speaker finishes his/her partner's utterance. What was not excluded in this study was that

speakers' gestures are aligned simply because they are talking about the same topic. It might well be that people who are given an identical production task (e.g. telling a story) always use the same gesture to express the same event. This issue was addressed in a follow-up study in which Kimbara (2008) used an experimental set-up. The author studied dyads jointly retelling a cartoon story, as was the case in the 2006 study, but this time there were two conditions. One in which the narrators were separated by a screen, and one in which they could see each other. In the former case there was significantly less gestural alignment than in the latter case. This proves that the alignment does not arise because the participants are performing the same task (and because they are describing the same events), but because they can either see or not see each other. In other words, speakers actually adapt their gestures to the gesture use of their conversational partners. Because the task setting of jointly retelling a story to a third participant is quite different from everyday speech, Holler & Wilkin (2011) repeated the visible vs non-visible condition to a more naturalistic dyadic setting. They asked dyads to play a card sorting game and either the participants had full visual access to their partner, or a screen prevented the participants from seeing each other. Similar to the results of Kimbara (2008), Holler & Wilkin (2011) found significantly more gestural alignment when participants were able to see each other compared to when they could not.

Drawing on the research sketched so far, Mol et al. (2012) addressed the question whether the observed gestural alignment is driven by processes of motor-mimicry (i.e. fully automatic and regardless of any semantics) or whether it is linked to meaning representation at the conceptual level. In two experiments the authors shed more light on this issue. In a first experiment, participants watched a video of someone telling a cartoon story. In two conditions the person in the video either performed *congruent* gestures (e.g. drawing a door when talking about a door) or *incongruent* gestures (e.g. drawing a door when talking about a dog). Participants who saw the video had to retell the story to a third participant themselves. In doing so, the participants performed significantly more aligned congruent than incongruent gestures. In fact only one gesture from the incongruent condition was repeated during the retelling. This indicates

that gestural alignment is not a mere formal copying, but that semantic representations play a role too. In a second experiment, Mol and colleagues found evidence that during a map task, participants not only adapt to their partners' gestures, but also adopt their perspective on the navigation task. In this experiment the confederates always used the exact same verbal description of a route on a map. In one condition, however, the confederate performed gestures indicating he conceptualised the navigation as a route on a vertically oriented map. In a second condition, the confederate used gestures indicative of a conceptualisation as a route through a city. Confederates and participants took turns in describing routes to each other. Participants systematically used the same type of gestures as their confederate partner. What this type of gestural alignment indicates is not only a formal repetition of gesture hand shapes, but also an adaptation to the conceptualisation expressed in the gestures of the confederates.

A somewhat different take on gestural alignment can be found in Parril & Kimbara (2006). These authors investigate how observing gestural (and verbal) alignment can affect subsequent gesture (and speech) behaviour. To address this issue Parrill & Kimbara had participants watch video clips of interacting dyads. Those video clips contained dyads either aligning or not aligning to each other. Afterwards the participants were asked to retell what they saw in the video clips. Participants who observed the aligning dyads reproduced more of the gesture features from the videos in their own retelling, which indicates a high degree of sensitivity to alignment as a phenomenon.

So far, all of the research on gestural alignment presented here used an experimental set-up of participants retelling stories, watching videos or playing games. A notable exception are Bergmann & Kopp (2012) who used a corpus (the SAGA-corpus, see Lücking et al. (2013) for details) of more spontaneous speech to study gestural alignment. Interesting from a methodological point of view is that, analogous to Howes, Healey & Purver (2010) for lexical and syntactic alignment (cf. supra), Bergmann & Kopp use *control dialogues* to test whether the alignment they measure occurs above chance level. A second relevant methodological issue is that Bergmann & Kopp (2012) did not code their gestures for one dimension

only. They annotated their data for five features: representation technique, handshape, palm orientation, finger orientation and handedness. Alignment was measured for each of those features separately. Bergmann & Kopp (2012) found significant gestural alignment but not for all the gesture features they coded. Only for the features representation technique, handshape, palm orientation and handedness the mean similarity of prime and target is higher than to be expected by chance. In line with what Gries (2005) and Howes, Healey & Purver (2010) found for lexical and syntactic alignment, Bergmann & Kopp (2012) demonstrate significantly more self-alignment than other-alignment. This effect was observed for all features coded for. A final important result in their work, was the relevance of *distance*: the more gestures between prime and target, the smaller the probability of prime and target being aligned. The effect of distance was found for the features representation technique and handshape. For the other features the alignment rates remain more or less constant with increasing distance.

In the present study we use "gesture" only to refer to bodily motion of arms and hands. However, for other types of bodily motion alignment has been observed as well. A key publication on research in this field is Chartrand & Bargh (1999). They coined the term *chameleon effect* to refer to the process of aligning postures, mannerisms and facial expressions during dialogue. The authors demonstrated the effect for a number of mannerisms: face touching, foot shaking and smiling. During a task of discussing photographs, participants were teamed up with a confederate that either performed the mannerisms under scrutiny, or did not. Compared to baseline data (recorded prior to the actual experiments), participants performed significantly more mannerisms if their partner (i.e. the confederate) also did. In a second experiment using the same task, the confederate was asked to either copy no or all of the mannerisms of the participant. A subsequent questionnaire revealed a link between alignment and liking: participants whose mannerisms were being copied by the confederate indicated the interaction ran more smoothly and they liked their partner more, compared to when they were not being copied. It is important to add that during the questionnaire the participants were unaware of the fact that they were being copied (or not). This adds proof to

the hypothesis that alignment, even in its most unconscious form, still has a social and communicative effect. A more recent study by Bailenson & Yee (2005) even reproduced this link between alignment (of mannerisms) and liking in conversations between humans and virtual agents. Shaik et al. (2013) show how the *chameleon effect* occurs in very young, 40 month old children, albeit without this social link. The children significantly copy more mannerisms such as yawning, frowning, face scratching and head wiggling when watching a confederate perform those actions than during a baseline. However, they do not copy the confederates more if they like them more (compared to confederates they like less).

The observation that conversational partners align their mannerisms, even in human-computer interaction or adult-child interaction, leaves undiscussed the issue of how this alignment might be synchronised. Shockley et al. (2003) and Louwerse et al. (2012) explicitly address the point of how non-conscious bodily behaviour is temporally organised across speakers (see *interpersonal synchrony* in Fig. 1). To be more precise, both studies use cross recurrence quantification analysis (CRQA, see section 4.1.2 in Chapter 4 for more details on this type of analysis) to dig into the matter. Shockley et al. (2003) used a magnetic body tracking system to measure participants' posture during a picture sorting task. What they found was that participants actually talking to each other displayed significantly more alignment of postural sway compared to participants who talked to a third, non-present confederate. Interestingly, this significant difference emerged regardless of whether the participants were able to see each other or not. In other words, there was no difference in alignment rates of postural sway in dyads that were facing each other, compared to dyads that were back-to-back. What did matter was the interaction between the dyads. Even if the dyads were standing face-to-face, but were each talking to a confederate outside the experiment room, there was no significant postural alignment. These results provide evidence for the hypothesis that alignment is not only based on motor mimicry, i.e. an immediate and formal input-output process, but also depends on conversational and semantic processing. Using the same analytical method of CRQA, Louwerse et al. (2012) add to these observations that synchronisation occurs not only for posture, but also for a wide range of

other dimensions of (bodily) behaviour. This study is particularly interesting because it combines alignment at the different levels discussed so far: lexical, syntactic and gestural (including bodily movements beyond hand and arm). Louwerse and colleagues show that synchronisation increases over time and over task difficulty. The longer participants talk to each other and the more cognitively demanding their task, the more they synchronise their aligned behaviour. Moreover, synchronisation also increases over a social dimension they controlled for. In the experimental set-up participants took turns in either giving or receiving instructions to navigate through a complex map. For most of the behaviours under scrutiny the instruction follower copied the instruction giver significantly more often than vice versa. This hints at the importance of conversational role in the directionality of alignment. Better informed or dominant conversational partners align less than their less informed, less dominant partners.

GAZE ALIGNMENT

In the multimodal array of levels at which alignment has been demonstrated, eye gaze holds a somewhat special position. This has to do with the dual function of gaze: it is both an instrument of perception and production. Of course, most of what our eyes do is perceive the world around us. However, the eyes can also convey communicative content. Bavelas et al. (2002) labelled gaze to be a *visible act of meaning.* Although gaze in itself hardly ever entails a symbolic form-meaning pairing, i.e. it does not carry propositional content, it does serve a multitude of social and communicative functions. This was already illustrated in the pioneering work of Kendon (1967) and Argyle & Cook (1976), but also more recent work focusses on those different functions such as managing attention (Langton et al. 2000, Vertegaal et al. 2001) and turn-taking (Cassell 1999, Novick 1996, Oertel 2012), providing feedback (Bavelas et al. 2002, Jokinen 2010), highlighting information structure (Cassell et al. 2000) or asserting social dominance (Kleinke 1986).

Although the observation that eye gaze can convey meaning is ancient[10], research on eye gaze and certainly on gaze alignment is very recent. Most

[10]Already Cicero (106-43 B.C.) is attributed the quote "Ut imago est animi voltus sic indices oculi": the face is a picture of the mind as the eyes are its interpreter.

studies on the topic start from a joint-attention paradigm in which participants look at external visual stimuli during a map, puzzle or matching game of some sort. Richardson & Dale (2005) for example, looked at the coupling of gaze behaviour in a separate production and perception task. First, a participant (the *storyteller*) was asked to talk about the cast members of the TV series *Friends*. During their storytelling they saw pictures of the main cast members on a screen in front of them. Afterwards, other participants (the *listeners*) had to listen to the speech of the storytellers. The listeners too were looking at a screen with the same pictures of the cast members. What Richardson & Dale (2005) found was that storytellers' and listeners' gaze was coupled. Using CRQA they demonstrated that typically two seconds after a storyteller looked at a picture the listener would look at the same cast member as well. In a follow-up study Richardson, Dale & Kirkham (2007) replicated their findings in an experimental set-up where participants actually spoke to each other (rather than *listeners* watching a video of *storytellers* in the 2005 study). They also added an experiment that illustrates how *common ground* (cf. Clark 1996) is linked to the coupling of eye gaze. In this experiment participants had to discuss a picture by the surrealist painter Dalí. Prior to the discussion, participants either received information on the content and the meaning of the painting or information on the life and theories of Dalí. Participants knew whether they received the same or different information. From the results it was clear that gaze alignment is linked to common ground between interlocutors: there was significantly more gaze alignment in the condition where participants received the same information (regardless of which information), compared to where they received different information.

Some researchers found an effect of increasing gaze alignment as discourse unfolds. Hadelich & Crocker (2006), for example, had people discuss pictures in a 3x3 grid on a screen in front of them (the interlocutors were unable to see each other). The authors showed that as the interaction

proceeded, the gaze alignment became tighter: the *eye-eye-span*[11] became shorter the longer people talked. Dale et al. (2011) replicated these findings in an experiment of participants playing a tangram matching task (again with interlocutors unable to see each other). They not only found that there was more gaze alignment towards the end of the experiment, but also that gaze alignment occurred faster the longer participants interacted. Moreover, the coupling of eye gaze appeared to be linked to task completion times: the more gaze alignment, the faster the participants solved the matching puzzle.

Some researchers have studied the influence of gaze on alignment at other levels. These studies might have been discussed in previous sections, but since they crucially involve a type of gaze alignment (i.c. participants looking at each other), we discuss them here. Postma et al. (2013) uncovered a link between gaze and intonation alignment. In a talkback experiment with virtual agents, participants were asked to repeat series of digits the virtual agent produced. Postma and colleagues provided evidence that when participants were being looked at by the agent, they aligned more (in terms of intonation) than when they observed the series of digits with averted gaze by the agent. Similarly, Wang et al. (2011) showed that gaze aversion plays a role in gestural alignment as well. Previous research (Heyes et al. 2005) had already shown that it is easier to perform a gesture (e.g. opening a hand) when primed with a video of that gesture than primed with a video of a different gesture (e.g. a hand closing into a fist). Wang et al. (2011) now found that it is also easier to perform a target gesture following a prime in which the participant is being looked at, compared to a prime without eye contact. However, this effect was only observed in the congruent trials, i.e. the trials in which the prime and target gesture were the same (both opening or both closing hands). In the incongruent trials, where participants were shown a closing hand but they

[11] Hadelich & Crocker (2006: 38) define the eye-eye span as "the time difference between the last fixation by the speaker to a referent before the onset of the respective referring expression and the first fixation by the listener to the same referent after the respective referring expression", i.e. the time difference between *gaze prime* and *gaze target*

had to perform an opening hand, gaze appeared to have no effect on the ease of performing the target gesture.

1.3.2 Alignment: a pervasive phenomenon serving multiple functions

The previous section (1.3.1) provided ample evidence that alignment is a pervasive phenomenon that occurs at many different levels. The fact that there is alignment in face-to-face (and even in human-computer or computer mediated) interaction is unquestionable. The question why people align is open to much more debate. In this section we will review some core attempts at answering that *why*-question.

COMMUNICATIVE FACTORS

Two communicative factors are linked to the question why people align. First, people align because it facilitates communication. Second, people align to achieve communicative goals. We start by zooming in on the former factor. Garrod & Pickering (2004) claim that "conversation is so easy" because interlocutors align at multiple levels of representation.

> "To come to a common understanding, interlocutors need to align their situation models, which are multi-dimensional representations containing information about space, time, causality, intentionality and currently relevant individuals. The success of conversations depends considerably on the extent to which the interlocutors represent the same elements within their situation models" (Garrod & Pickering 2004: 8).

As already pointed out in the section on lexical alignment (cf. supra) these authors advocate an automatic view on alignment that is driven by mechanical priming effects. When perceiving a prime, listeners activate the mental representations (or *situation models*) that go with that word (or phoneme, or construction, or intonation or whatever). This increased activation makes it more likely, and also more efficient, to use that same word (or phoneme, etc.) in a subsequent utterance. In doing so, conversational partners not only align their linguistic production but crucially also their mental representations to that of their conversational partner, which is what successful communication is all about.

Pickering & Garrod's theory implies that both comprehension and production efficiency benefit from alignment. This claim has been backed-up by some experimental research on the topic. Even before the introduction of Pickering & Garrod's interactive alignment theory, Chartrand & Bargh (1999) showed that interlocutors who are being copied (in terms of mannerisms, cf. supra) indicate their interaction ran smoother compared to interlocutors that were systematically not copied by a confederate. This increased smoothness was also reported in follow up studies by Chartrand, Maddux & Lakin (2005) and Van Baaren et al. (2009).

Alignment is not only linked to conversation smoothness, but also to task performance. In Reitter, Moore & Keller (2006) and Louwerse et al. (2012) we see that the more difficult a task gets, the more interlocutors align at different verbal and non-verbal levels. Also, Nenkova, Gravano & Hirschberg (2008) found that alignment of high frequency words is a good predictor for task completion times. Although we should be cautious and not take correlation for causation, for example Porzel, Scheffler & Malaka (2006) provide further evidence for a causal link between alignment and task performance. In an experiment using the Wizard-of-Oz paradigm[12], a confederate either lexically copied the participant or explicitly refrained from copying him/her. Participants were asked to install and use a new digital television. What the authors found was that when participants were lexically copied by the confederate, they were significantly faster in completing the task. However, there was a cross-effect of technical expertise: lexical alignment only resulted in faster task completion for non-expert participants who were new to the whole concept of digital television. More technically well-versed participants who were used to working with digital television did not benefit from the alignment effect.

So far, we have only described a first communicative factor in answering the question why people align. Apart from the communication-facilitating function of alignment described above, interlocutors also align to achieve specific communicative goals. Turning to less automatic and more conscious forms of alignment, and turning to less experimental and

[12] Within this type of research in human-computer interaction (HCI), participants are asked to communicate with what they believe is a computer. In fact, it is a human confederate, which allows for certain types of control over the interaction.

more CA (conversation analytic) methods, Tannen (1987), for example, showed how alignment functions as a conversation management tool. Speakers linguistically align with their partners to claim the floor, signal listenership or provide back-channelling cues. Also within the CA tradition, Perrin et al. (2003) demonstrate how alignment (i.c. *diaphonic repetition*[13]) functions as a means of signalling attention or (dis)agreement. In this vein interlocutors are shown to re-use immediate linguistic content to indicate they are still listening or they (dis)agree with what their partner just said. Research within a cognitive linguistics tradition (see the section on syntactic alignment and *dialogic syntax* above) further demonstrates how interlocutors use alignment to achieve personal communicative goals. For example, speakers construe their utterances parallel to that of their conversational partners to exploit the rhetorical potential of alignment and to trump their conversational partners (Zima 2013).

In the context of interaction between native speakers (L1) and second language speakers (L2), Costa et al. (2008) found that alignment is also used to signal understanding. This indicating that you have understood your partner occurred both from the part of the L1 and the L2 speaker. Moreover, the authors show how alignment causes L2-L2 conversations to be easier than L1-L2 conversations. Compared to mixed dyads, pairs of L2 speakers share the same mistakes and non-native construals and constructions (especially if their mother tongue is the same). This sharing of non-native linguistic forms involves a higher level of alignment at the representational level as well: to them the automatic priming and alignment of mental representations works just fine. In mixed L1-L2 dyads, a non-native prime would not be recognised by the L1 speaker, forcing him to a cognitively costly, non-automatic modelling of what his L2 partner is trying to say (and vice versa for the L2 speaker not understanding all the native linguistic input). This makes for less easy and less smooth conversations, compared to the L2-L2 cases.

[13] Perrin and colleagues (2003: 1844) distinguish *diaphonic repetition* from other types of repetition in that it "clearly manifests the speaker's intention to quote and therefore reproduce and qualify what the interlocutor has just said. As a form of quotation, such repetitions indicate an individual's intention to comment upon the other's talk, by taking it up as the theme and object of reference of his own discourse".

Overall, there seems to be sufficient evidence in favour of the communicative function of alignment: conversations with more alignment are (perceived as) running smoother and are more likely to lead to better task performance. Moreover, interlocutors use alignment in a multitude of conversational strategies.

SOCIAL FACTORS

People align because they want to belong to or distance themselves from a group. Alignment is also shown to correlate with how (positive) we assess others. Recurrent in many studies on social factors underpinning alignment is the link between liking and alignment. This link works in two ways: participants like others more if those others align to them, and participants align more to others they like. Evidence of the former is given in Chartrand & Bargh (1999): participants like confederates with whom they talked more if those confederates copied their mannerisms (compared to when those confederates did not). This result was replicated by Bailenson & Yee (2005) for interaction between humans and computers: embodied agents that copied participants' head movements were judged to be more persuasive and likable than agents performing pre-recorded head movements. Not only for mannerisms, but also for prosodic features the link between liking and alignment seems to hold true. Nass & Lee (2001) for example, show how participants who are copied in terms of voice intensity, pitch and speech rate like their conversational partner (i.c. a computer) more than in the condition without the copying. This link between liking and alignment even appears to exist if the alignment is of a very abstract, non-communicative nature: Hove & Risen (2009) demonstrated how it works for tapping fingers on a drum computer. In their experiment, participants were asked to watch a digital metronome and tap their finger on a drum computer in sync with that metronome. In three conditions a confederate in the same room would either tap the drum computer synchronous to the participant, asynchronous to the participant or not tap at all. In a subsequent likeability test, the confederates in the *synchronous* condition score significantly higher than those in the asynchronous or no tap condition. Crucial here is that the effect resides in the synchrony rather than in the alignment, viz. it was not enough for the confederate to copy

the behaviour of the participant (i.c. the tapping): the alignment also needed to be synchronised in order to resort an effect on the likeability test.

That participants like confederates more if the latter aligned more to the participants, is clear from the studies listed above. The effect alignment has on participants, however, goes beyond liking. People who are being copied prove to show more prosocial behaviour in general than their counterparts who were not copied. Van Baaren et al. (2004, 2009) provide convincing evidence of this. They found that participants who are being copied in terms of mannerisms by a confederate, are more helpful towards that confederate. For example, when at the end of the conversation the confederate drops his pens, participants who had been copied were significantly more likely to help pick those pens up, compared to the non-copied participants. This effect even transferred to other confederates: also new confederates (confederate 2) who only entered the room after the conversation between the participant and confederate 1 were more often helped by participants who were copied than those not copied. The effect is shown to transfer even further, in the sense that when asked to donate part of their financial compensation for taking part in the experiment to charity, the copied participants give significantly more than the non-copied participants.

We have just demonstrated that people are more prosocial or people rate partners as more likeable if their conversational partners copy them. However, the other way around, people also copy partners they like more. Stel et al. (2010) demonstrate this for facial expressions and mannerisms. In an experiment they ask participants to watch a video of a confederate talking into the camera. All participants get to see the same video, but prior to seeing it they are given different background information on the person in the video. In one condition Stel and colleagues elicit negative liking, in another positive liking towards the confederate in the video. The results show that participants, although they saw the exact same video, aligned significantly more in terms of facial expressions and mannerisms to the confederate in the video in the positive liking condition compared to the negative liking condition. Similarly, but also reversely, people align less to conversational partners they do not like. In a fairly early

study on the topic, Bourhis & Giles (1977) demonstrated this for speech accent. They showed how Welsh participants start talking with a heavier Welsh accent when they are talking to an arrogant English interviewer (a confederate), compared to talking to a neutral interviewer. The ability of speakers to align or dis-align from their conversational partners when they like or dislike the partner in particular is striking. In this vein, linguistic or behavioural distance appears to serve as a proxy for social distance. Alignment and liking seem to mutually feed into each other.

People not only align more to people they like, also other beliefs than liking shape speakers' alignment behaviour. We have already described the research on alignment by Costa et al. (2008) in terms of communicative efficiency. L1 speakers adapt their verbal behaviour when talking to L2 speakers to facilitate communication. This *foreigner talk* is driven by the beliefs of the L1 speaker that the L2 speaker lacks the necessary proficiency to fully understand the native L1 language. In other words, L1 speakers align to the non-native verbal behaviour of L2 speakers not because they like them but because they believe them to be less proficient. This effect has also been demonstrated in human computer interaction. Branigan et al. (2004) found that participants align more (both lexically and syntactically) to what they believe to be a computer than to what they believe to be a human. In a follow-up study (Pearson et al. 2006), the authors show that even very subtle information shapes these beliefs. They asked participants to interact with a computer but used two conditions. One in which the participants were led to believe the computer was very old and basic, and another where the computer was presented as state-of-the-art and advanced. Although in fact participants were interacting with the same computer, they judged the *advanced* system to be more competent. More interestingly, participants lexically aligned significantly more to the *basic* computer than to the *advanced* one. This increase in alignment observed by Pearson and colleagues boils down to the same principles behind the results in Costa et al. (2008): speakers adapt their alignment rate to how proficient they believe their partner to be.

Speakers not only adapt their alignment behaviour to their beliefs about their conversational partners, but also to beliefs about themselves. This was demonstrated by Uldall et al. (in prep.) who asked participants to

complete a personality test. Regardless of what the participants filled in, they were either told that their personality was very close to average, or very different from average. In a subsequent conversation with a confederate the authors found significantly more lexical alignment in the *different* than in the *average* condition. Apparently, participants copied the confederate more if they believed themselves to be special or at least divergent from average. This ties in with Brewer's (1991) theory of *optimal distinctiveness* which claims that people balance on a desire towards distinctiveness versus assimilation, i.e. people want to be unique and different (but not too much), but at the same time they want to be similar to the people around them (but not too much). Alignment, it seems, is a mechanism that is susceptible to this balance: participants who just heard their personality is quite off, compensated this imbalance towards distinctiveness by acting more like the next person they talked to (i.c. the confederate).

NEUROLOGICAL/BIOLOGICAL FACTORS

People align because they are hardwired to do so. Also, alignment is said to have evolutionary roots. Among others, Lakin & Chartrand (2003) and Dijksterhuis & Bargh (2001) argue that automatically aligning to certain behaviour by fellow humans is relevant to survival. For example, seeing other people run away without aligning to that perceived action, might cause life-threatening situations. Crucial in their account is the automaticity in alignment, i.e. the direct link between perception and behaviour. Thanks to the mechanics of priming, there is no need to further model or interpret the other's behaviour. This absence of cognitive effort speeds up the time between perception and behaviour, which can be crucial in terms of survival situations. Furthermore, since human beings are per definition social beings, it is important to affiliate to or distance ourselves from social groups when needed. From the previous section it was clear that alignment can both be the result of or the reason for positive social interactions. In other words, it has a strong regulatory function in social behaviour and therefore alignment is judged to be quintessential to survival in the inherently social world we live in.

Evidence from the hardwiredness of alignment comes from neurological research on, amongst other topics, mirror neurons. For example Iacoboni et al. (1999) and Rizzollati et al. (2001) have demonstrated that perceiving hand movements (e.g. of a hand grasping a cup) activates the same brain regions than performing those hand gestures. This effect has not only been demonstrated for object-directed hand movements but also for communicative gestures (Montgomery, Isenberg & Haxby 2007). If during the perception of an action by speaker 1 the brain area for the production of that action in speaker 2 gets activated, this neurologically facilitates and renders more likely an aligned action by speaker 2. Mirror neurons thus seem to be good candidates to underpin automatic priming, i.e. directly linking perception to production, and hence to underpin alignment. However, the discovery of the mirror neuron system is not enough to fully account for how alignment works. In the first place, and referring back to Fig. 1 at the beginning of this chapter, not all instances of alignment can be due to automatic priming. Humorous retorts in which conversational partners draw on each other's verbal behaviour cannot be fully explained by the intricate mirror neuron system. Second, recent research has shown that mirror neurons can be *overruled* by for example training (Catmur et al. 2007). Newman-Norlund et al. (2007) showed how the mirror system is more active during complementary action than during imitative action. They observed more mirror neuron activity in participants who were shown a cup that was reaching out for them to grab (i.e. complementary action) than when seeing someone grab a cup (i.e. imitative action). In this sense the mirror neuron system is not a rigid input-output system, but it appears to be flexible and adaptable. This is counterevidence for the assumption that only mirror neurons can explain alignment and that mirror neurons can only explain alignment.

1.3.3 Positioning & research questions

In the past two decades, both in psychology and linguistics, there has been a shift from studying (resp. the mind or the language use of) people in isolation to studying people in group. Because, increasingly, in both disciplines there is research on human *interaction*, rather than on individual cognitive or linguistic processes, we have witnessed a shift in unit of

analysis. No longer the individual participant, but the dyad (or group) constitutes the main unit of analysis. This becomes clear from comparing older interaction models to their more recent counterparts. Human interaction is no longer perceived as an alternation of speakers and listeners coding and decoding information (cf. Jakobson 1960: 353 or Shannon 1948: 381), but more as a complex adaptive system of interlocutors that mutually constrain each other's interactional contributions (Beckner et al. 2009, Fusaroli et al. 2014, Tollefsen & Dale 2012). In this vein, conversational partners are not regarded as individual cognizers each contributing their personal share to the ongoing interaction, but as a joint system of contributions to a *shared* project. As a consequence, rather than the other way around, interaction shapes language, grammar or even meaning and cognition as such.

With this dissertation we want to add to the growing body of research on interaction, and more specifically we will contribute to the study of how alignment shapes interaction. Most of the literature on alignment reported here can be situated in the domain of (cognitive) psychology, however, we want to add evidence and insights from a (cognitive) linguistics perspective. This means we are primarily interested in conversational processes, rather than cognitive processes. We adopt an interactional linguistic point of view and do not concern ourselves explicitly with explaining *why* people align, but rather refining *when* and *how* people align[14]. This might seem like taking a step back, after all, *that* interlocutors align during interaction has long been established. However, a better understanding of when and how interlocutors align is useful. For example, if alignment appears to be linked to social factors such as liking, knowing when people align might predict whether people like each other. In most studies this measuring of alignment happened very fragmented (only looking at one feature of facial expressions), coarse-grained (one specific construction to stand for syntactic alignment in general) and monomodal (not looking at the interplay between multimodal layers). How can we develop a more fine-grained method of measuring linguistic alignment? And how does linguistic alignment relate to other types of alignment? If

[14] Of course, by digging into the temporal and multimodal dimension of alignment (cf. infra), we do contribute to a better understanding of the *why* question.

people align lexically to their partner, can we see the same thing happening at the gestural level? Or is gaze alignment (i.c. mutual gaze) a good predictor for gestural or lexical alignment? Or for both? And do the same predictive factors such as distance between prime and target, relative frequency of the prime or characteristics of the person performing the prime apply to alignment at different levels?

An issue that has not been addressed systematically in alignment research is the question *when* people align. Does alignment get established slowly and evenly throughout a conversation? Or does it occur abruptly at a crucial point in the interaction? Or are there even more complex temporal dynamics to it? To answer these questions, and to structure the questions themselves before we try and answer them, we further position the present work in the literature overview presented above and then formulate our exact research questions.

TEMPORAL DIMENSION

So far, we have reported studies that demonstrate the existence and prominence of alignment as a phenomenon, and studies that provide reasons *why* people align during interaction. With the notable exception of Louwerse et al. (2012) and some studies on gaze alignment (Hadelich & Crocker 2006, Dale et al. 2011) hardly any attention has been paid to the temporal dynamics of alignment. Apart from the exceptions stated above, nearly all the studies on alignment measure whether one group of participants aligns more than another group. This approach makes complete abstraction of how alignment might not only differ between groups but also between different points in time within one group. In this dissertation we want to explicitly focus on the *temporal dimension* of alignment. Interlocutors' goals, attention, feelings towards each other, etc. change throughout the course of an interaction. If this is the case, than we expect alignment to do the same thing. What studies such as Louwerse et al. (2012) have shown is that during a goal-oriented task interlocutors synchronise their aligned behaviour more (at different levels) the longer they interact. What we do not know is whether this temporal pattern is due to the task itself. Do we see the same pattern for different tasks? Or for conversations where there is no task at all? The research questions for the

temporal dimension of alignment can be summarised and structured as follows:

(i) Do interlocutors align more the longer they interact?
(ii) Or do we see a different temporal pattern for alignment?
(iii) Are the temporal dynamics of alignment task dependant?

Parallel to obtaining answers to these research questions we try to achieve two methodological subgoals as well. First, we want to develop methods to measure and map the temporal dynamics of alignment. These methods should be transferrable to other conversational phenomena than alignment. What the method will entail, is an analysis of alignment at different levels of *granularity*. At the speech level for example, it is possible to look for recurrences ranging from character bigrams to complete clause constructions. Some of the levels in between, e.g. morphological units, have not yet been considered. Crucially, we want to test what the implications are of measuring alignment at different levels of granularity. The same goes for different levels of *abstraction*: it is possible to look for linguistic alignment at token or type level, or at the level of POS-tags (part-of-speech), or at the most abstract level of checking whether there is alignment or synchronisation of the mere presence of speech.

A second methodological subgoal is our endeavour to help research on alignment *into the wild*. So far, most of the studies start from the safe haven of the lab in which a lot of control over the interactions is possible. Studying less controlled types of interaction does not have these experimental advantages, but can contribute to the observations and theory building on alignment in other ways. By studying a combination of different interaction types we hope to methodologically contribute to studying alignment in more naturalistic settings.

MULTIMODAL DIMENSION

Research on alignment is multimodal in the sense that alignment at different semiotic channels has been observed. What is lacking, to date, is research on the interaction and trade-off between those multimodal levels. In this dissertation we will look at how gaze, gesture and speech alignment

relate to each other. More specifically we want to answer the following research questions:

(i) Do the same factors predict alignment at different levels?
(ii) Does alignment at one correlate with (dis)alignment at another?

Of course, the multimodal dimension presented here, should not be regarded as independent from the temporal dimension. In fact, we want to blend the research questions and arrive at a *multimodal* analysis of the *temporal* dynamics of alignment (see Fig. 3). If we study the impact of alignment at one level on alignment at another level, we will also incorporate the temporal dimension and ask ourselves whether this impact is constant throughout time, or whether the multimodal interplay is temporally dynamic itself. Also, when trying to find which factors are good predictors of alignment at different modalities, we will factor in temporal dimensions as well. Similarly, when we study whether alignment gradually increases throughout a conversation, and whether such a temporal pattern is dependent on the type of task participants perform, these issues will be addressed from a multimodal perspective, i.e. we will study the temporal dynamics of alignment not only for speech but also for gesture and gaze.

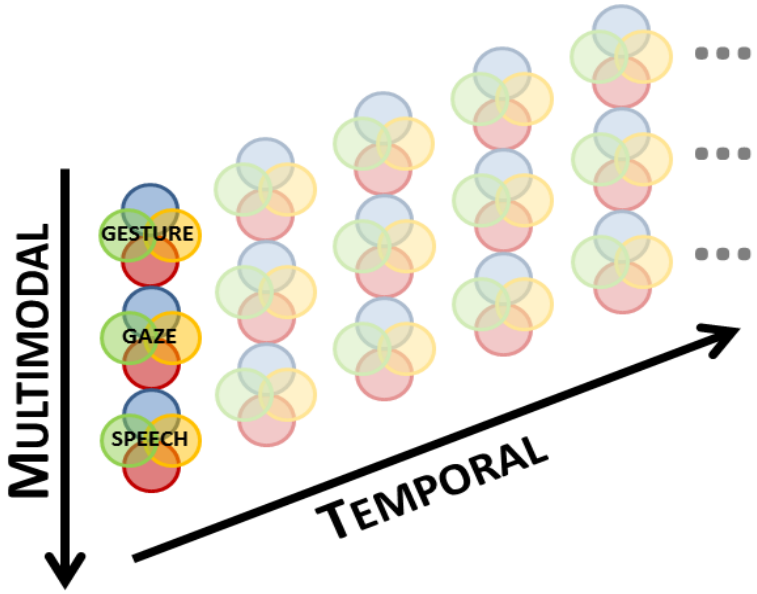

Fig. 3: Integrating the temporal and multimodal dimension

Finally, as is indicated in Fig. 3 (and referring back to Fig. 1) we want to apply our integration of the temporal and multimodal dimension to different types of phenomena: we will study the temporal dynamics of multimodal alignment not only for intentional but also for unintentional behaviour, not only for semantically rich but also semantically poor behaviour, not only for concrete tokens but also for abstract types, and not only for individual words, but also for more complex constructions. Taking this step away from answering the question *why* people align, we will provide a lot of evidence on the questions *how* and *when* people align. This in itself can help shed light on the underpinnings, be it cognitive or conversational, of alignment, and thus contribute to the ongoing debate on the foundations of this pervasive phenomenon.

Chapter 2

The corpus

2.1 Introduction: corpus requirements

To systematically study the research questions sketched out in the previous chapter, we need a corpus with a specific set of characteristics. First, we need a corpus of spoken interaction. Because alignment is an interactional phenomenon by nature, most written corpora and spoken corpora of monologues are unsuited. What is more, the dialogue type has to be sufficiently spontaneous and undirected: scripted telephone calls or news studio interviews are often well prepared and too rehearsed to allow for a 'natural' measurement of alignment. We do not want our participants to use the word "vehicle" for the concept CAR because it is written on an autocue or read from a protocol paper. We want their linguistic (and non-linguistic) behaviour to be unbound and interactionally motivated.

Second, audio recordings do not suffice for our purposes. Central to this study is the multimodal approach to the phenomenon of alignment. Apart from (and in combination with) linguistic behaviour we also want to demonstrate how alignment works at the gestural level. This means we need video recordings in which all gestures are visible and recording settings in which gesturing is not restricted. This latter point means we want to avoid studying gestural behaviour in situations where participants are using their hands for eating, writing, holding microphones, etc.

Third, because we want to factor in eye gaze as a predictor for gestural alignment and because we want to study gaze-speech and mutual gaze synchronisation, it is imperative that we accurately measure what interlocutors are looking at during their conversations.

Finally, to maximally address our research questions we have to be able to compare across subjects and/or objects. This will involve balancing between experimental control and spontaneity. Studying free range conversations can be very instructive, but a more experimental approach is required. In what follows, we first give an overview of some existing corpora that might be up to the job of serving as data set (2.2), to then demonstrate why and how we compiled our own corpus (2.3).

2.2 Existing corpora

Capturing and annotating video data is a time-consuming task. Guided by technological advances and driven by a shared interest in multimodal

dialogue, the past decade or so, researchers in many different subdomains have devoted their time to compiling multimodal dialogue corpora. We have no ambition to provide an exhaustive overview (see Allwood 2008, Knight 2011 or Paggio et al. 2013 for good overviews), however, we do want to single out some corpora to illustrate how they do and how they do not fulfil the requirements listed in the previous section. What is lacking to maximally be able to answer our research questions, can be related to two main issues: the camera set-up and the measuring of eye gaze.

2.2.1 The issue of camera set-up

Crucial in developing video corpora is the recording set-up: conversations can be recorded with one, two or multiple cameras. The issue with using one camera is that the resulting image yields a static external perspective on the interaction: speakers are captured either in profile (as in Gerwing & Allison 2009, McNeill 2008 or Pine et al. 2004) or facing the camera (as in Bertrand et al. 2008 or Kimbara 2006). Fig. 4a, a still taken from the CID Corpus of Interactional Data for French (Bertrand et al. 2008), and Fig. 4b, a still taken from Gerwing & Allison (2009), illustrate that measuring gesture and gaze behaviour based on single-camera perspectives is problematic either way. A profile shot too often blocks the most distant arm and hand, and facial information from view (Fig. 4b), whereas a frontal shot creates an unnatural angle at which the speakers are forced to interact (Fig. 4a).

Fig. 4a: Frontal shot in the CID corpus

Fig. 4b: Profile shot in Gerwing & Allison

These issues can be overcome by using multiple cameras. Several corpus projects have adopted a multi-angle approach, with either a speaker- or a scene-oriented focus. In a speaker-oriented design, the primary focus is on

capturing the individual speakers in as much detail as possible. In a scene-oriented setting, the cameras are set up in such a way that the analyst obtains a 360° perspective on the interactional landscape (or an approximation thereof).

Examples of speaker-oriented multi-angle corpora are the NMMC Nottingham Multimodal Corpus (Knight et al. 2008, see also Fig. 5a) and the IFA Dialogue Video Corpus (Van Son et al. 2008, see also Fig. 5b). In both corpora the interactions were recorded using two cameras positioned next to the speakers and facing the other. Although this allows for a more natural face-to-face setting and a more reliable coding of facial features, still not all the requirements set out in section 2.1 are met. The most important remaining issue is the lack of a full view on the entire gesture space. Most notably in the set-up of the IFADV corpus, a lot of the gestures are lost because the camera is zoomed in too much on the face and upper torso. The camera perspective in both Fig. 5a and 5b is double, but it remains static: if participants move, they move out of the viewing frame of the cameras.

Fig. 5a: Double camera perspective in the Nottingham Multimodal Corpus

Fig. 5b: Double camera perspective in the IFADV corpus

A scene-oriented recording technique was adopted in the D64 corpus (Campbell 2008), the VACE multimodal meeting corpus (Chen et al. 2006), the NOMCO corpus (Paggio et al. 2010) and the UTEP-ICT Cross-Cultural Multiparty Multimodal Dialogue Corpus (Herrera et al. 2010). In all four examples, two- or multiparty interactions were recorded using multiple cameras (up to ten in the case of the VACE corpus). In Fig. 6a-b we show the recording layout for the D64 and VACE corpus. The benefit of this type of set-up is that speakers are never out of sight of the cameras: every gesture is recorded and can be assessed in more detail because it is available from multiple perspectives. The annotation of eye gaze remains problematic (see

2.2.2), but the multiple camera recording technique allows for a better estimation of eye gaze compared to single camera perspectives. A further drawback is the massive amount of data and the correspondingly massive amount of annotation time needed to process and analyse these data.

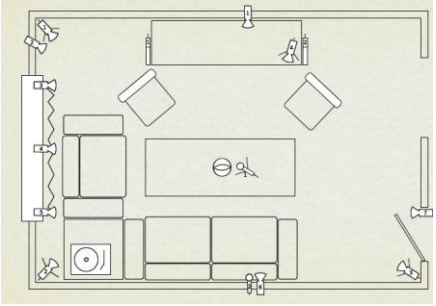

Fig. 6a: Recording configuration for the D64 corpus

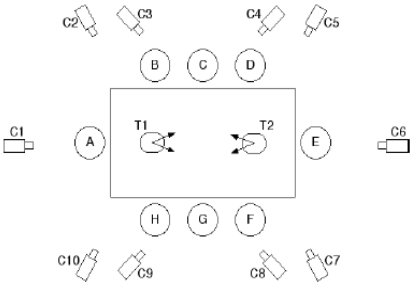

Fig. 6b: Recording configuration for the VACE corpus

Recently, a number of researchers have explored the possibilities to reduce this processing and annotation time by harvesting online videos for (semi-) automatically creating multimodal corpora. They use closed captions (i.e. subtitles, texts for the hearing impaired, textual overlays, graph captions, etc.) to annotate and index those video files. One example is the UCLA NewsScape Archive (also known as the *Little Red Hen Database*), which contains more than 200,000 hours of television and video news programs from 2005 to the present, indexed by three billion words of closed captioning, transcripts, and on-screen text. The collection is accessible (although not publicly) through an online search engine at http://newsscape.library.ucla.edu/. Another example is the TV NEWS Archive (free of charge and publicly accessible at https://archive.org/details/tv). This online corpus contains more than 624,000 news programs collected from 2009 onwards. Both corpora are updated with new broadcasts continuously and older materials are being added as well.

These large scale video corpora provide high quality and -most importantly- a gargantuan amount of video data. However, there are two major drawbacks for these corpora to be well-suited for the present study.

First, although there are many cameras recording the interaction from different perspectives, only one perspective at a time occurs in the broadcasted edit. Because directors switch between camera perspectives (close-ups, audience images, graph inserts, switching between interviewer and interviewee, etc.) speakers are not in the video continuously. This makes that a lot of valuable information is lost. A second drawback concerns the interaction type: because most of the data are television shows (and more specifically news, sports or weather reports) they are highly scripted and not dialogical.

2.2.2 The issue of measuring eye gaze

One dimension of multimodal speaker behaviour that is included in several of the above-mentioned corpora, but which poses a significant challenge in the corpus annotation process, is eye gaze. Studying gaze behaviour on the basis of video data of a participant's face, as was done in the VACE corpus (see Fig. 6b) and the IFADV corpus (see Fig. 5b), is problematic. This issue is explicitly raised and acknowledged by -among others- Kendon (2004), Paggio et al. (2010: 2970) and Streeck (2009: 117-118).

One way to overcome this problem of reliability and accuracy is to track participants' gaze behaviour using eye-tracking equipment. Recent research has used eye-tracking in two types of studies: (i) experimental studies in which participants look at external visual stimuli during a map, puzzle or matching game of some sort (Brennan et al. 2008, Dale et al. 2011, Frischen et al. 2007, Lachat et al. 2012, Neider et al. 2010, Richardson & Dale 2005), and (ii) interactional studies where interlocutors can see each other during their face-to-face conversation (Al Moubayed et al. 2013, Jokinen 2010).

Although all of the examples in (i) and (ii) accurately measure the eye gaze of the interacting participants, none of them are suited for the present study. In (i) speakers are separated from each other, either by a wall or by having them communicate via computer screens (as in Fig. 7). The researchers in (i) study on which target objects -on screen or physically present in the lab- participants focus. They are not concerned with how speakers look at each other, or how they look at common events in a shared physical space. Moreover, the experimental set-up and the task

design are so specific to the individual studies, that sharing the video data is problematic. The videos are intended as private data sets, rather than open corpora.

Fig. 7: Still from the recording set-up in Brennan et al. (2008): eye-trackers measure what speakers look at on their displays, but interlocutors are physically separated

The examples in (ii) do have speakers engaged in physical, face-to-face interaction, but the issues here are that not all speakers are being eye-tracked, or that there are a lot of restrictions regarding spontaneous gestures. Fig. 8 illustrates how the recording setting can obstruct interlocutors in gesturing freely. Even if they would gesture, the set-up of the cameras does not allow for a sufficient view on all participants' gestures. Moreover, the use of table top eye-trackers in (ii) requires the participants to sit still and keep their heads within the viewing range of the eye-tracking system. Again, this is not an ideal basis for a conversation with spontaneous co-speech gestures.

Fig. 8: Still from the recording set-up in Al Moubayed et al. (2013): eye-trackers measure fixations, but measuring spontaneous co-speech gestures is problematic

2.3 Creating the Insight Interaction Corpus

From the previous section it is clear that using an existing corpus would not allow us to answer all of the research questions outlined in the first chapter. Although time-consuming, compiling a tailor-made corpus to perfectly fit our research needs was needed. We built a multi-camera video corpus with participants engaged in different types of interaction, during which their eye gaze was measured with an eye-tracking device. The multiple cameras allowed for a detailed view on all of the gestures, the eye trackers provided accurate gaze information, and the task design elicited both spontaneous conversations as well as experimentally controlled data with repeated lexical and gestural references to predefined target items. The resulting Insight Interaction Corpus (Brône & Oben 2015) is not only of use for the present study, it is also available for other researchers interested in topics on multimodal interaction (see corpus project page with samples at https://www.arts.kuleuven.be/ling/midi/corpora-tools/insight-interaction-corpus). This section will further elaborate on the set-up, task design and annotation of the Insight Interaction Corpus.

2.3.1 Recording set-up & devices

RECORDING ROOM

A screenshot of the recording configuration is presented in Fig. 9. The pairs of participants sit opposite each other with no objects within their reach (this to avoid parts of the body being concealed on the video images and to maximize the freedom for hand gestures). Behind each of the subjects is a large screen on which animations are projected during a collaborative task (see section 2.3.2). A fixed camera records the ongoing conversation in profile (left image in Fig. 9), and two cameras integrated in the head-mounted eye-trackers provide a frontal view on each of the conversational partners (right images in Fig. 9). The green dots[15] on the right images are generated by the eye-trackers and indicate exactly what the participants

[15] The mobile eye-trackers provide two types of data: video files from a scene-camera, and data files on the basis of the eye-movements (containing simple x and y co-ordinates that together constitute the exact location of the fixation point, at a rate of 30 Hz). The right images in Fig. 6 are an overlay of the video files from the scene-camera with the gaze co-ordinates from the data files.

are visually fixating. In the example in Fig. 9 one of the girls is looking at her partner's face, the other is looking at her partner's gesture. On the basis of the external camera perspective alone (left image in Fig. 9) it would be impossible to reliably discriminate between fixating the face or fixating the hand gesture.

Although there is a wire connecting the eye-trackers to the computers, the subjects are free to move and gesticulate. They do not need to restrict themselves to a certain position or virtual frame. The cameras on the eye-tracking glasses are flexible (see full red circle in Fig. 10a) , which allowed us to adjust them in such a way that the subjects saw their partners' eyes at all times. However, at no point (not even during the calibration of the eye-trackers) did the participants see their own or their partner's eye gaze behaviour, i.e. they never got to see the eye-tracking software (see Fig. 10b), showing a close-up of the eye or the green dot moving about (as in the rightmost images in Fig. 9).

Fig. 9: Recording configuration in the Insight Interaction Corpus

RECORDING DEVICES

Our recording set-up allowed us to obtain video data from three perspectives (fixed camera + two scene cameras), eye-movement data for each of the participants (two eye-trackers) as well as the audio signals (two microphones). Below is a list with some of the technical specifications of the gear used during the compilation of the corpus.

- 1 fixed colour camera
    - Sony HDR-FX1000E
    - 25 frames per second
    - 720x576 pixels
- 2 head-mounted eye-trackers, with scene camera included
    - Arrington Gig-E60 eye-tracking glasses
    - 30 frames/events per second
    - 320x240 pixels
- 2 microphones
    - Zoom H2 : directed
    - Both microphones recorded in the 16bit/44.1kHz WAV format

Fig. 10 further shows how the head-mounted eye-trackers work. Each of the eye-tracking glasses is equipped with two cameras. The scene camera (see dotted red circles in Fig. 10) records what is in the viewing field of the person wearing the glasses. The eye-tracking camera (see full red circles in Fig. 10) captures the position of the subject's pupil. The software (Fig. 10b) then maps the information from the two cameras onto each other: by superimposing a green dot over the video images from the scene camera, we know what the subjects are visually focussing on within their viewing field.

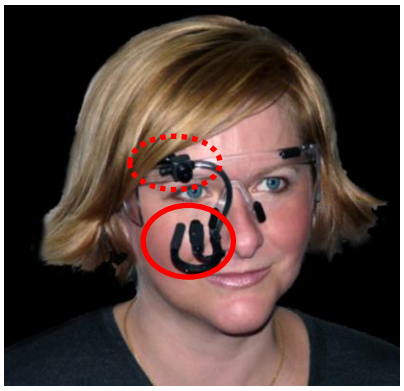

Fig. 10a: The eye-tracking glasses with a scene camera (dotted line) and an eye-tracking camera (full line)

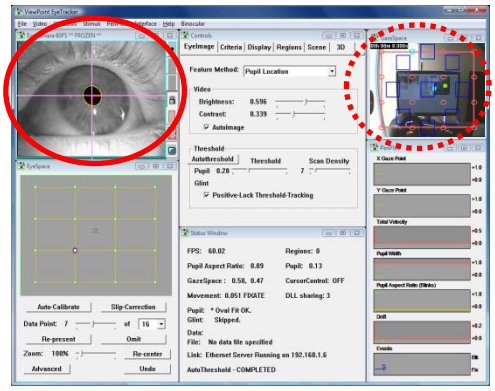

Fig. 10b: The eye-tracking software mapping information from the eye-tracking camera (full line) onto the scene camera image (dotted line)

2.3.2 Task design

When put in a (recording) room together, the last thing two people will do, is remain absolutely silent. To get participants to talk to each other does not require clever experimental design. However, we wanted to be able to compare across subjects and to create baseline conditions (see section 3.2.2 in Chapter 3). To enable that, we needed different people talking and gesturing about the same things. Once the participants were all geared up in the lab, we had them engage in three types of task-based interactions: a cartoon description task, a problem solving task and a brainstorm task. We will only discuss the latter two tasks, since the first was only intended as a distractor for the subjects to get used to the experimental conditions, and as a technical buffer for the researchers to tackle problems with the eye-tracking, recording or projection devices.

ANIMATION DESCRIPTION

The animation description task is similar to the *diapix* used by Van Engen et al. (2010), in which participants play 'spot-the-difference' games on the basis of complex drawings. In that study, the subjects were asked to identify the differences in the drawings they were both shown (they could not see each other's pictures). In the present study, subjects had to find out differences between short video animations. Below we present two such

animations and briefly explain the differences between the animation of participant 1 and participant 2. First, consider Fig. 11a-b: the animation for participant 1 (Fig. 11a) starts with a square with a black dot hovering above it. Next, the dot follows the path indicated by the dotted blue lines: it moves below the square and turns red, moves to the right of the square remaining red and travels to the left of the square turning green. Whenever the dot passes the square it disappears 'behind' it. The difference with the animation for participant 2 (Fig. 11b) is that the coloured dots are outline in black and that the dot ends top left (compared to centre left in Fig. 11a)

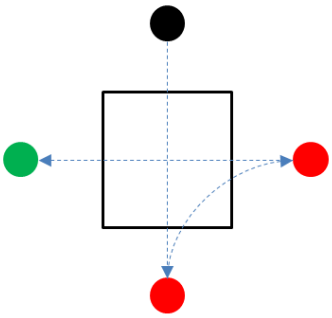

Fig. 11a: Example animation for participant 1

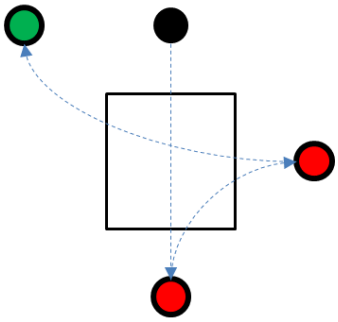

Fig. 11b: Example animation for participant 2

A second example can be found in Fig. 12a-b. What we see in Fig. 12 is the animation for participant 1. The animation starts with two lampposts: one with a dog behind it and the other with a cat in front of it (Fig. 12a). Next, the dog starts peeing and the cat starts circling around the right lamppost. Finally, a huge boot appears, moves from top to bottom and squashes the cat and the dog (Fig. 12b). In the animation for participant 2 the dog was not peeing and the boot did not move from top to bottom but from left to right. Some of the animations were more complex than the illustrations we presented here, but an elaborate description of all the video animations would lead us too far. Crucial is that the target objects in the animations (e.g. the cat and dog; the dot and the square) allow for cross-speaker comparisons: by making different speakers linguistically and gesturally refer to the target objects, we can measure how and how often they align in doing so.

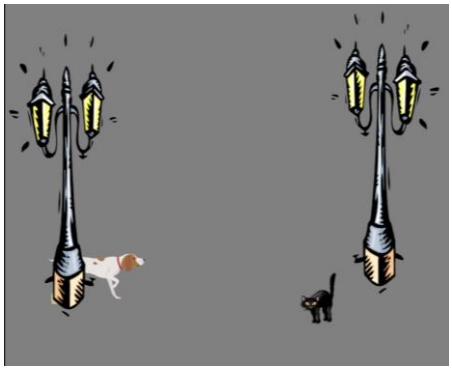

Fig. 12a: Example animation for participant 1

Fig. 12b: Example animation for participant 2

BRAINSTORMING

Next to the animation description task, we recorded the participants as they engaged in a less controlled type of interaction. By not restricting ourselves to one interaction type we were able to rule out that the alignment effects we measure are dependent of the specific conversational task. In other words, it allowed us to control that our results are at least generalizable beyond one specific experimental lab setting. To get the conversation in this part of the experiment going, we gave the speakers a brainstorming task. They had to come up with innovative or fun (physical and functional) features for a new smartphone, specifically designed for women. The results of some pre-tests revealed that participants relate to this topic and enjoy talking and fantasising about it. The participants were entirely free how rigorously they stuck to the given topic and how much time they spent on the task.

2.3.3 Procedure

Thirty participants (9 male, 21 female) from the University of Leuven (Campus Leuven and Campus Antwerp) took part in the study. All were native speakers of Dutch and between 19 and 31 years old (M=21.96, SD=2.58). Participants signed up in pairs that knew each other well prior to the start of the experiment. They received financial compensation for their participation and gave informed written consent to use the video

recordings for scientific purposes. The high level of acquaintance made them share a lot of personal and conversational history, which given the unnatural lab setting, made them more relaxed and less intimidated by the experimental circumstances.

As a distractor and to let the participants grow familiar with the lab setting, we asked them to watch a short cartoon and retell the story to their partner (who did not see the cartoon). After both participants had seen and retold their story, the eye-trackers were recalibrated and the actual experiment started.

First, participants were presented with short video animations ranging from abstract geometric shapes (triangles, squares, stars moving over, under and around each other) to more life-like situations (cars driving by on a road, sharks chasing divers, people hiding in a house, etc.). Each animation lasted ten to fifteen seconds. Participants were shown each video animation at the same time and could not see each other's animation. After having watched the animation, participants' screens went blank and they were asked to discuss the animation and discover the difference(s) between the animations. After each discussion, a new set of animations appeared, with a total of 15 animations. The total discussion time was about 15 minutes per dyad (M=15.24, SD=2.55). Every five animations, the experimenters entered the room to check the calibration of the eye-trackers, dividing the experiment into three blocks of five animations.

After the 15 animations, the researchers again checked the calibration of the eye-trackers and briefed the participants on the brainstorming task. The researchers only re-entered the recording room when the participants signalled they wanted to end the task (this varied between 4 and 12 minutes, M=6.20, SD=2.29). The total study lasted for about thirty minutes, excluding the initial calibration of the eye-trackers.

This procedure provided data that meets all of the requirements stated in section 2.1. As a result, the Insight Interaction Corpus data allow for a multimodal approach to the phenomenon of alignment in face-to-face interaction. Although the conversations are not fully spontaneous, there are sufficient arguments to categorize them as 'naturally occurring':

- The tasks create an external trigger to communicate and even impose a conversational topic, but the internal dynamics of the conversation is entirely free.
- There is no intermediate person or factor interfering with the ongoing discourse: the participants decide for themselves how long they talk, what their strategy is to perform the tasks, what they talk about apart from the ongoing task, etc.
- After a few minutes in the lab setting participants seemed relaxed, not paying attention to the recording devices and started chatting and joking cordially about topics totally unrelated to the experiment.

2.3.4 Synchronisation

When working with time-aligned data coming from five different sources, synchronisation is crucial. Because we only needed to cross-check the waveforms, synchronising the fixed camera with the microphones was straightforward. We used the editing tool Adobe Premiere Pro to perform this first synchronisation: it simply sufficed to load the video file from the fixed perspective and the audio files from the microphones, play them at the same time and listen whether there was any echo or an even larger difference between the three files. In none of the cases further manual editing was needed: each of the recordings ran 100 % in sync for the entire duration of the recording session.

Synchronising the video data from the eye-trackers was more time-consuming because there were a number of dropped frames and no exact information on where those frames were dropped[16]. To avoid a post-processing frame-by-frame analysis, we adopted a more practicable methodology: we looked for at least two anchor points per minute at which

[16] The average length of the video files was 6.36 minutes. The average number of dropped frames per video file was 38, with nearly all of the dropped frames occurring in clusters of 3 to 7 frames.

we checked the synchronisation between the three video files[17]. The data from the fixed camera was regarded as our fixed starting point or baseline and was left unchanged. If frames were dropped in the eye-tracking data, the corresponding span between two anchor points was 'stretched' so as to synchronise with the span in the baseline. This way we made sure that the dropped frames did not accumulate into a noticeable and undesirable time lapse between the video signals.

2.3.5 Annotation

For the transcription and annotation of the video data we used the audio and video annotation software ELAN (Lausberg & Sloetjes 2009, see also http://tla.mpi.nl/tools/tla-tools/elan/). Additionally, we used Praat (Boersma & Weenink 2009) because this tool is fully compatible with ELAN and visualises the amplitude and pitch of the acoustic signal. This was particularly helpful for the annotation of accent and intonation contour (cf. infra). In what follows we zoom in on the different levels the video corpus was annotated for.

TRANSCRIPTION

To transcribe all speech in the video data we used the GAT transcription norm (Selting et al. 2009). The advantage of that norm is its modularity, i.e. there are various possibilities to personalize the norm and choose the granularity or detail at which the acoustic signal can be represented. For our corpus we use an orthographical transcription with the following additional prosodic information:

- the main accent per intonation unit
- terminal intonation contour per intonation unit
- pauses within intonation units
- manifest lengthening of vowels and consonants

[17] On average we had an anchor point in the video files from the eye-trackers every 21.38 seconds. The exact number and position of the anchor points depended on the content of the video data: the onset or offset of hand gestures were particularly frequently used as anchor points because those actions were clear signals in each of the video files.

GESTURE

Inspired by the coding schemes for gesture by Allwood et al. (2008) and McNeill (1992, 2005), part of the Insight Interaction Corpus is annotated for gesture in very much detail (see Oben & Brône 2015 for the extended coding scheme): for five of the interactions in the corpus we annotated for gesture features such as handedness, handshape, palm orientation, finger orientation, etc. In this dissertation, however, we approach the phenomenon of gestural alignment from a holistic perspective, i.e. we very generally want to inquire into the question whether two gestures are the same (see also the discussion in section 3.1.2). Therefore, we coded the entire Insight Interaction Corpus for the feature of *representation technique*. To be more precise, each depictive gesture, viz. a gesture representing an object or an action, was coded using the typology suggested by Streeck (2008: 292-295). Streeck distinguishes gestural depiction methods such as modelling (hand as a token for an object), bounding (hands indicate sides or edges of an object), drawing (fingers draw lines that represent the outline or path of an object), handling (hands enact a prototypical usage of the represented object), etc. To check for the consistency of this gesture annotation we calculated kappa values. These values provide a statistical measure of inter-coder agreement, and are more robust than percent agreement (i.e. the relative number of labels that two annotators share) because they take into account the difference between observed and expected agreement. For example, if two annotators have to categorise participants' hair colour into "light" and "dark", they have a 50% chance of choosing the same label as their co-annotator. However, if they have to discriminate between 20 possible shades of hair colours, they will not easily choose the same label by chance. Therefore, a percent agreement of 70% might be more indicative of the low number of possible categories than of a high inter-coder agreement. The kappa statistic takes this issue into account. For the annotation of representation technique in the Insight Interaction Corpus, three coders annotated a random selection of 70 gestures. We calculated three kappa scores because each annotator is compared to each other annotator and we found values of 0.864, 0.751 and 0.770, which indicates good to very good coding consistency.

POS-TAGGING

For the part-of-speech tagging in this corpus we used the Frog tagger (Van den Bosch et al. 2007). Because the Frog algorithm was used for the reference corpus of spoken Dutch (CGN, *Corpus Gesproken Nederlands*, Oostdijk 2000), it is well trained for interactional data in Dutch. We ran the tagger with a reliability threshold of .06: if the Frog tagger attributed a POS-tag with 94% or more reliability, the tag was not checked further. Tags below that threshold were checked by a human annotator. To check for the validity and consistency of the tags attributed by the human annotators, we performed a kappa-test on a random set of 300 tags for 2 annotators (κ=0.892), which turned out to be satisfactory.

GAZE

When annotating gaze, the first important issue to be addressed is the minimum fixation duration. Gaze behaviour often seems very chaotic, as people switch from one object in focus to the other extremely fast. In other words, before starting the analysis of what people look at in the course of an interaction, we need a clear definition of what is considered as 'looking at', i.e. a standard of minimum fixation duration that allows for a reliable categorisation of a gaze event as a fixation. In the extensive literature on the topic, the most frequently used standard for minimal fixation duration is around 120 ms (Jacob & Karn 2003, Vertegaal et al. 2001). Although the eye-tracking device allows for a higher frequency, we used the frame rate of the video data (25 frames per second) to define the smallest possible unit of analysis, i.e. 1/25 seconds or 40 ms. Hence, in our annotation of gaze, participants need to focus on an object for longer than 3 frames (120 ms) before we recognised it as a gaze event. Fixations of 3 frames or less were disregarded as relevant units.

Although we used a mobile eye-tracking system that allowed the participants to move, the corpus only contains video data from dyads sitting opposite each other. This restricts the number of potentially relevant objects or regions that can be fixated. As a consequence, the tag set for gaze contained a limited set of items: face (of other), body (of other), gesture of other, own gesture, own body, screen, floor, ceiling and misc.

The gaze behaviour of each of the participants is annotated for the entire duration of the video file.

The annotations of the eye-tracking data were checked for consistency, using the same technique as described in the section on gesture above. We checked a total of 70 randomly selected gaze events, and found kappa values of 0.947, 0.902 and 0.911. These near perfect values show a more than sufficient coding consistency at the gaze level.

2.3.6 Privacy

All of the participants in the corpus have signed a document in which they agree that the recordings can be used for scientific research. The document explicitly mentions that excerpts of the audio or video files can be used in publications or presentations. Furthermore the corpus has been declared at the Belgian federal privacy committee CBPL (Commissie voor de Bescherming van de Persoonlijke Levenssfeer). All of the video and audio files have been anonymised in the sense that references to existing people have been omitted in the media data and replaced with fictional references in the transcriptions.

2.4 Conclusion

Although multimodal corpora are growing in quantity and quality, creating our own data set was necessary. With the Insight Interaction Corpus presented in this chapter we now have a video corpus of unscripted, face-to-face interactions. The recording set-up maximally allows participants to gesture freely, with those gestures being recorded from different camera perspectives. Not only gesture, but also eye gaze is reliably mapped thanks to the mobile eye-trackers. The task design of animation descriptions and brainstorming is of course specific to the topic of alignment in this study, but by making the corpus available we hope that the donkey work of annotating the video files will be of benefit to other researchers as well.

Chapter 3

Exploring the multimodal dimension

In this chapter we take on a multimodal perspective towards alignment. In two case studies we demonstrate (i) how gaze behaviour and (ii) how a range of contextual factors are able to predict lexical and gestural alignment. These studies are multimodal in the sense that we check whether the same factors hold for different modalities (i.c. gesture and speech), and whether one semiotic channel (i.c. gaze) affects another (i.c. gesture or speech).

3.1 Case study 1: Gaze as a predictor for lexical and gestural alignment

In the early works on the role of eye gaze in interactional settings (mainly Argyle & Cook 1976 and Kendon 1967) the focus was not on the *interactional* coupling of eye gaze, but on the *intrapersonal* coupling of it. For example, it was studied how eye gaze correlates with one's own speech, but not how it correlates with the speech or gaze of the conversational partner. More recently however, researchers are investigating interactional aspects of gaze behaviour. A first topic in this respect is that of *shared gaze*. One of the basic features of interaction is the joint focus of attention of co-participants. One correlate of this basic feature is shared gaze, i.e. the joint visual focus on relevant aspects of the context (e.g. referents that are the current topic of conversation). This type of gaze alignment in which interlocutors adapt their gaze behaviour to that of their partner has been discussed in several recent studies (Dale et al. 2011; Hadelich & Crocker 2006; Richardson & Dale 2005; Richardson, Dale & Kirkham 2007). The research question in these studies typically boils down to 'do interlocutors look at the same thing at the same time?'. The rate and the speed with which interlocutors in a joint task focus on the same referents has been shown to correlate with task performance (Brennan et al. 2008, Neider et al. 2010), shared knowledge (Richardson et al. 2007, 2009) and duration of the conversation (Hadelich & Crocker 2006, Dale et al. 2011).

A second topic in research on interactional aspects of gaze behaviour (as opposed to intrapersonal aspects), is that of *gaze cueing*. This can be defined as the effect that cueing a target (i.c. by looking at it) has on the gaze behaviour of an addressee. It is, put colloquially, the fact that

looking at something makes other people look at that same thing too. The crucial difference between gaze cueing and the shared gaze described in the previous paragraph, is that gaze cueing can only occur if interlocutors can see each other. In all of the studies on shared gaze described above, interlocutors were separated from each other. Their shared gaze is a residue of their verbal interaction. Interlocutors are looking at the objects because they are talking about them. Crucial to gaze cueing is that interlocutors are fixating an object, not because of the verbal interaction, but because their partner is fixating that object. Studies on the gaze cueing effect, which date back to early work by Posner et al. (1980), stress its role for joint attention in interaction (Emery 2000, Frischen et al. 2007). However, these studies still perform their experiments in a non-face-to-face setting in which participants are presented with a photograph or picture of a conversational partner. What these researchers found, is that participants are faster (in terms of reaction time) at targeting an object if that object was first fixated by the pair of eyes in the photograph or picture. Lachat et al. (2012) are the first to test this type of experiment in a face-to-face setting (rather than in on-screen experiments), however without using eye-tracking to obtain gaze data. Gullberg & Holmqvist (2006) and Gullberg & Kita (2009) focus on one specific case of gaze cueing, using head-mounted eye-trackers, viz. the effect a speaker fixating his own gesture has on the addressee's gaze behaviour. Both studies reveal that a speaker's gaze at own gestures is a powerful cue for addressees to leave the dominant fixation position (i.e. the face of the speaker) and give overt visual attention to the speaker's gesture.

A third and final relevant interactional aspect of gaze is the effect known as the *audience effect*. Basically, people behave differently when they know or have the feeling they are being watched. More specifically, studies such as Bateson et al. (2006), Piazza, Bering, & Ingram (2011) and Powell et al. (2012) have demonstrated that people behave more empathically and pro-socially when they are being watched. For instance, they pay more for their drinks, cheat less and donate more to charity, compared to when they are not being watched. This effect is even shown to hold for situations in which no actual human eyes, but (schematically

drawn) pictures of eyes on a piece of paper constitute the eye gaze stimulus.

The above-mentioned studies show how eye gaze has an effect on subsequent behaviour: fixating an object makes other people fixate that object too (*gaze cueing)* and being looked at affects the amount of money people put in a box (*audience effect*). However, while the studies reported so far do address interactional aspects of gaze behaviour and the influence of gaze on subsequent behaviour, none of them show how gaze plays a role in *alignment* at other semiotic channels. In this section we will investigate whether eye gaze can enhance lexical alignment (3.1.1) and gestural alignment (3.1.2). To be more precise, in terms of gestural alignment, does it matter whether or not a prime gesture has been focussed on by the speaker, by the addressee or by both (cf. *gaze cueing*)? In terms of lexical alignment, does it matter whether or not speaker and addressee are looking at each other during the utterance of the prime and/or the utterance of the target (cf. *shared gaze*)?

3.1.1 Gaze and lexical alignment

Knowing what people are visually focussing on provides us with information on people's attention, but also on people's mental and emotional states (Jones et al. 2006, Langton et al. 2000, Mathews et al. 2003). In the present study we will focus on the former aspect of attentional states. In digging into the relation between gaze behaviour and speech behaviour we hypothesise that gaze direction influences the degree of verbal alignment. More precisely, if interlocutors fixate each other's face, we expect them to lexically align to each other more frequently. This hypothesis is driven by related studies by, for example, Postma et al. (2013) and Wang et al. (2011, 2014) who found comparable gaze effects for alignment of intonation and hand shape respectively. Chartrand & Bargh (1999) also provide evidence for our hypothesis, but they framed it the other way around: if interlocutors are aligning, they look at each other more often. In one of their experiments a confederate either deliberately mimicked the participants' mannerisms, or he did not. Chartrand & Bargh report more eye contact in the mimicking conditions, compared to the non-mimicking conditions, which further exemplifies a link between eye gaze and alignment.

What -to the best of our knowledge- has not yet been addressed, is the relation between gaze and lexical alignment. Compared to related studies on the interplay between eye gaze and alignment, this case study is discerned by the interactional setting in which we study the phenomenon. Wang et al. (2011, 2014) and Postma et al. (2013) study gaze in a non-communicative setting: participants get to see video images (of either an actor or an avatar) and are explicitly asked to perform a certain type of behaviour. For Wang and colleagues that is producing simple target gestures (viz. opening or closing the hand), for Postma and colleagues it is reproducing simple words (viz. digits between 0 and 10). For both researchers the experimental conditions are twofold: the actor or avatar in the video is either looking at the participant (SpeakerGaze+) or not (SpeakerGaze-). However, what separates the two, is the method of measuring the dependent variable *alignment*. Whereas Wang et al. (2011, 2014) measure reaction times, viz. between the prime gesture in the video and the target gesture performed by the participant, Postma et al. (2013) measure to what extent there is intonational alignment, viz. in terms of pitch between the prime word in the video and the replication of that word by the participant. In the present case study the dependent variable is lexical alignment, viz. whether or not participants use the same word to refer to the same object. This is crucially different in that participants in our data set are actually talking to each other, face-to-face, and are not instructed to (re)produce certain gestures or words. Although they are performing a task, they are completely free in whether or not, how often, and how they label the target objects under scrutiny.

A second difference with the studies on the interplay between gaze and alignment reported so far, is that they only consider *shared gaze* (or the absence thereof) during the prime, when factoring in eye gaze. However, in dyadic interactions, more gaze configurations are possible (see Table 1 in the methods section below). In our data set, both during the utterance of prime and target word, interlocutors are entirely free to fixate whatever they want. In the experimentally controlled studies reported above, the addressee is instructed to always look at the video, and the actor or avatar in the video always performs the same gaze behaviour during the prime as during the target.

Third, and overarching the previous two issues, is the fact that this case study taps into *functional* and *communicative* aspects of eye gaze, next to *cognitive* or *social* aspects. Because both gaze and alignment behaviour are not elicited or controlled for in our data set, participants will have different reasons to fixate (or not) their partner, and to align (or not) to that partner. However, if participants look at a video in which an actor or avatar is performing a simple gesture or word, and they are instructed to perform a given or a matching gesture or word, those participants have no communicative or functional reasons for their behaviour. Hence, the effect gaze has on the subsequent behaviour can only be cognitively or socially[18] motivated, not communicatively or functionally. We argue that in real, face-to-face interaction, those communicative functions matter as well. To better understand this, we will not only study shared gaze during the prime, but also disentangle the gaze behaviour between speaker and addressee, and prime and target.

Following up on what Chartrand & Bargh (1999) did for mannerisms, we will first study whether shared gaze, viz. eye contact in which interlocutors are fixating each other, correlates with more lexical alignment. Second, and compared to the studies by Postma et al. (2013) and Wang et al. (2011, 2014), we are not only investigating the gaze behaviour of the speaker during the production of the prime, but also at the gaze behaviour of the addressee, and at the gaze behaviour during the utterance of the target. Postma and Wang and colleagues already found that being looked at (i.e. SpeakerGaze during the prime) correlates with higher or faster alignment rates. This study will add to that observation whether or not fixating the speaker (i.e. AddresseeGaze, either during prime or target) affects alignment rates as well.

The added value of the present study is not as much that we check the effect of gaze for yet another multimodal layer (i.c. lexical alignment), but that we use eye-tracking in a spontaneous face-to-face setting to obtain more fine-grained gaze data in more naturalistic interaction. In doing so, we are not only looking into the cognitive or social underpinnings for the

[18] We interpret *socially* in this case as relating to the audience effect described above: participants' behaviour is affected because they know they are being watched.

observed alignment behaviour, but also into the functional and communicative ones.

METHOD

For this case study we use the Insight Interaction Corpus described in Chapter 2. However, we only consider the second interaction type, i.e. the animation description task. In this task, the interlocutors were each shown a video animation. They saw the animation at the same time, but they could not see each other's animation. The two video animations were identical, except for a few minor details. The goal of the task was to discover those differing details. Once they completed the task, they were shown a new animation (with a total of fifteen animations they had to discuss). For more details on the data set and its annotation, see Chapter 2.

ANALYSIS

To quantify gaze behaviour, we use the annotation from the Insight Interaction Corpus (see 2.3.5). In the corpus, all fixations on the face of the conversational partner are annotated as such and there is no further differentiation between fixations on the eyes, nose or mouth of the partner.

In quantifying lexical alignment we basically want to measure whether or not participants use the same words to refer to the same target objects they are discussing. To measure this type of lexical alignment we take interactional prime-target pairs as our unit of analysis. To illustrate how we define such pairs, consider the example below. In this example the participants are discussing a video animation in which a cat and a dog are performing the actions (cf. section 2.3.2 and the animation described in Fig. 12).

S1 First there was a **cat** and a dog.
S2 It was a black **cat**.
Did you have a black **cat** as well?
S1 Yeah.
S2 Well they started, the dog was like peeing all the time.
[…]

S2 And the **pussy** was circling, I guess it was clockwise, was circling round and round a lantern post.

S1 In my case the **pussy** was circling, the **pussy** was, I don't know, clockwise or, no I don't remember. But very fast anyway. I couldn't count how many times.

S2 The **pussy** was smaller than the dog?

We define prime-target pairs as adjacent lexical references to the target objects in the animation videos that are produced by different speakers. In Fig. 13 below, the prime-target pairs for the example above are schematically represented and marked in green rectangles. The second ("cat") and third ("cat") lexical reference in the example are adjacent, but they are produced by the same speaker. Therefore they are not a prime-target pair. Similarly, lexical item one ("cat") and three ("cat") do not constitute a prime-target pair either: although they are produced by different speakers, they are not adjacent.

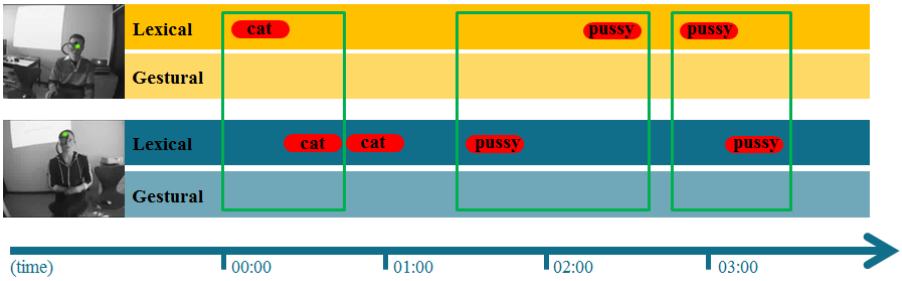

Fig. 13: Lexical references to the target object CAT in which all of the prime-target pairs are aligned

To annotate lexical alignment, we identified all the prime-target pairs in the corpus (n=723) and digitally scored them for alignment: either speakers use the same word or they do not. In the example above, there are three pairs (see green rectangles in Fig. 13) that are all aligned. Although annotating lexical alignment is a digital matter, the two lexical items in the prime-target pairs need not necessarily be fully identical in order to be counted as aligned. For example, we discarded diminutives and plurals and regarded cases of "katten" (cats) and "katje" (little cat) as identical and thus fully

aligned to the root form “kat” (cat). Only in cases where the root forms in the interactional pair differed (like in “kat” (cat) vs. “poes” (pussy)) we considered the items in the pair as not aligned.

Lexical alignment will be the dependent factor in this case study; gaze the independent. We have already explained how we measure gaze behaviour in terms of fixations on the face, but because of the dyadic conversational situation, four levels of gaze behaviour are relevant:

(i) Gaze behaviour of the **speaker** during the **prime**
(ii) Gaze behaviour of the **speaker** during the **target**
(iii) Gaze behaviour of the **addressee** during the **prime**
(iv) Gaze behaviour of the **addressee** during the **target**

At either of these levels the gaze can be directed towards the face of the interlocutor or away from the face[19]. In the example above, all three prime-target pairs are aligned, and there always is eye contact, except for the third pair. In this case the speaker is looking away when he utters the prime, and the addressee is looking away when listening to the target word. Next to gaze behaviour at these four levels, we also calculated *eye contact*, both during prime and target. Consequently, there is eye contact in the first two pairs and no eye contact in the third pair. Our scoring table for this simple example, then, would look like this:

[19] In fact, in some cases (n=19) a gaze shift occurs during the utterance of a prime or target lexical item. Because we only want to include cases in which there either is a full fixation or a full gaze aversion during the entire duration of the lexical item, we deleted case of gaze shifts during the utterance of a lexical item from the data set.

	Alignment	(i) Speak Prime	(ii) Speak Target	(iii) Address Prime	(iv) Address Target	Eye Contact Prime	Eye Contact Target
pair1	1	1	1	1	1	1	1
pair2	1	1	1	1	1	1	1
pair3	1	0	1	1	0	0	0
…	…	…	…	…	…	…	…

Table 1: Example coding scheme of resp. the dependent factor alignment, the independent gaze factors ((i)-(iv)), and the calculated factors for eye contact during prime and target

Using mixed effects models we first want to uncover whether eye contact during prime or target enhances lexical alignment rates. Second, we look at gaze behaviour in a more fine-grained way, and try to demonstrate how gaze behaviour of the speaker (focussing on the face of the addressee) or the gaze behaviour of the addressee (focussing on the speaker's face) during the prime or during the target are good predictors of whether or not the lexical items in prime and target are aligned. We use mixed effects models[20] because we maximally want to take into account idiosyncratic alignment behaviour. Some dyads may systematically align all the time, or some dyads may produce much more prime-target pairs than others. The same goes for the target objects that might each typically favour or disfavour lexical alignment. To avoid taking variation in alignment rates that is due to specific dyads/objects for variation due to our independent factor of gaze behaviour, we treat *dyads* and *objects* as a random factor in our model. Furthermore, to test that our independent variables are truly independent, we calculated Cramer's V measures for every possible interaction between the independent variables. None of those measures exceeded 0.16, which provides evidence against such a collinearity issue where one independent factor is too good a proxy for another.

[20] We computed all of the mixed effect models in this dissertation using the LME4 package for R, described in Baayen (2008: 263-308). To obtain p-values, we used the LMERTEST package.

RESULTS

Our research question was whether gaze behaviour during face-to-face interaction affects lexical alignment. A first relevant result in answering that question comes from zooming in on the effect of *eye contact*, i.e. cases where both speaker and addressee are looking at each other's face. Our results indicate that the average alignment rate is higher in cases of eye contact, both during the prime and during the target (Fig. 14). If there is eye contact during the prime, in 91.8% of the cases prime and target are aligned (compared to 71.3% without eye contact). Similarly, if eye contact occurs during the target there is more lexical alignment (89.9%) than when there is no eye contact (80.9%). A mixed effects model, with *dyad* and *object* as random factors (cf. supra), lexical alignment as dependent factor, and eye contact during prime and target as independent factors, shows that the difference in alignment rates between the presence and the absence of eye contact is significant only for the prime ($z=2.641$, $p=0.008$) and not for the target ($z=0.871$, $p=0.38$). There was no interaction between the two. This means that if both addressee and speaker are looking at each other during the utterance of the prime, chances of the prime-target pair to be aligned are significantly higher. Whether or not there is eye contact during the target, seems to be of little importance in terms of lexical alignment

The mixed effects model just described appears to be sufficiently explanatory. A first indication thereof is the C-value for the model ($C=0.86$), suggesting (near to) predictive power. Second, we compared the fitted value for each data point to the actual value in the response variable[21] and found that the model predicted 89.6% of the data correctly. This a considerable improvement compared to a naïve model, i.e. the average score of lexical alignment for the entire data set (86.8%). Although this is already a very high baseline to top, the mixed effects model still scores notably better.

[21] We rewrote the fitted values into a binomial dataset, with fitted values larger than 0.5 as predicting alignment (value "1"), and smaller than 0.5 predicting absence of alignment (value "0").

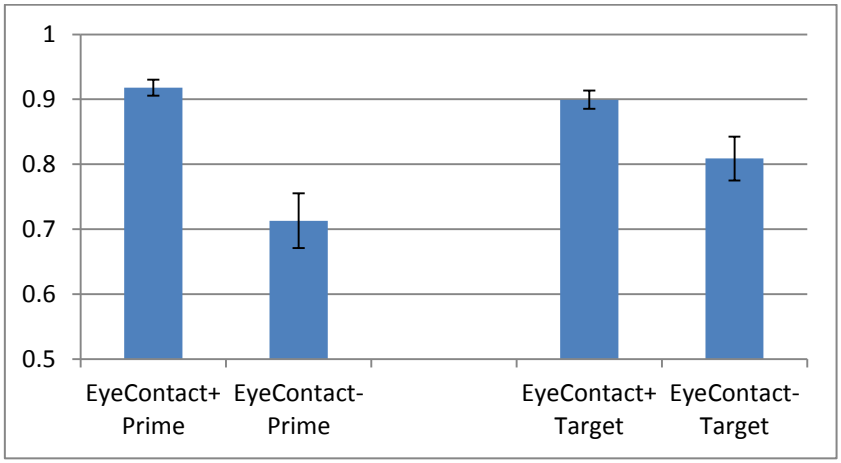

Fig. 14: Average lexical alignment rates for cases of eye contact and no eye contact during the utterance of the prime and target. Error bars indicate standard error.

So far, we have only presented results concerning *eye contact.* However, our data allow for a more fine-grained analysis. We have already established that eye contact is relevant in predicting lexical alignment, but only during the prime. What we also find is that, more precisely, the behaviour of the speaker during the prime is key. A mixed effects model with dyad and object as random factors, lexical alignment as dependent factor and the factors ((i)-(iv)) concerning gaze behaviour described in table 2 as independent factors, shows that the gaze behaviour of the speaker during the prime is the best predictor ($z=5.074$, $p<0.001$). No other factors and no interaction between factors reached significance. This means that when a speaker is fixating the addressee while uttering the prime, the addressee will be more likely to align to that prime (alignment rate of 91.5%, see Fig. 15), compared to when the speaker averts his gaze during the prime (alignment rate of 66.7%, see Fig. 15). The model for this result had a satisfactory C-value of 0.86, and when comparing the fitted to the actual data points, the model was correct in 89.6% of the cases (see footnote 21). Given the high performance of a naïve model based on the average score for lexical alignment alone (86.8%), the mixed effects model adds relevant explanatory power.

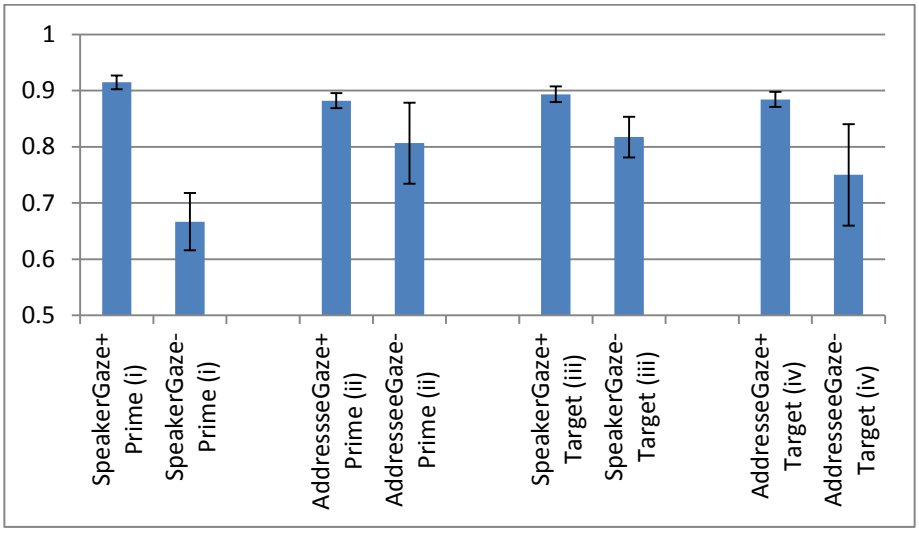

Fig. 15: Average lexical alignment rates for gaze levels (i)-(iv). Error bars indicate standard error.

DISCUSSION

What this case study shows is that gaze behaviour during the prime, and not during the target, significantly affects lexical alignment. If there is eye contact during the utterance of the prime, we observe more lexical alignment. To a large extent, this effect can be explained by the gaze behaviour of the speaker alone: if the speaker fixates the addressee during the prime, that addressee will significantly more often use the same lexical reference during the target, compared to when the speaker does not fixate the addressee during the prime.

These results tie in with previous research on the interplay between gaze and alignment of mannerisms (Chartrand & Bargh 1999), gesture (Wang et al. 2011) and intonation (Postma et al. 2013). Eye contact, and more specifically SpeakerGaze during the prime, correlates with higher alignment scores. This effect seems to hold true not only under controlled lab settings in which video stimuli of conversational partners are used (Postma et al. 2013, Wang et al. 2011) but also in naturalistic face-to-face interaction. This is an important finding given the wide variety of functions gaze has, and given the very idiosyncratic nature of gaze behaviour during conversations. As reported by Kendon (1967) or Cummins (2012), there is a

lot of between-subject variation in gaze behaviour. For example, although participants fixate their partner more during listening than during speaking, different studies (Argyle & Ingham 1972, Cummins 2012, Kendon 1967, Nielsen 1962) report more variability within the categories "gaze during speech" and "gaze during listening" than between those categories. Some dyads hardly look at each other and other dyads look at each other constantly. Because of this large between-subject variation, the amount of looking at a partner's face is a bad predictor for speakership or listenership. In other words, eye gaze in this example is more dependent on who is talking (between-subjects), rather than whether participants are talking (between-categories). What the results in this case study indicate, is that regardless of this strong speaker-tied variation of gaze during face-to-face interaction, speaker gaze is a good predictor for lexical alignment.

A surplus of this study compared to related work, is that we not only account for the gaze behaviour of the speaker during the prime, but also for addressee gaze, and also for gaze behaviour during the target. This allows us to discriminate between the perception and the production role of gaze. What our results suggest, is that for lexical alignment, the production side of gaze matters more than the perception side. Fixating the speaker (AddresseeGaze) during the prime can be linked to perception, viz. to increased attention to what the speaker is saying. This does not enhance lexical alignment. Fixating the addressee during the prime (SpeakerGaze) can be linked to the production side of gaze, viz. to a signalling function: by looking at his addressee the speaker can highlight that something relevant is going on in his speech signal. This type of gaze behaviour does enhance alignment rates at the lexical level. Related to this issue are the baselines of gaze during speech and during listening. Kendon (1967), Argyle & Cook (1976) or more recently Cummins (2012) found that speakers typically look away while speaking, and listeners typically look at the speaker while listening. Therefore, SpeakerGaze (i.e. speaker fixating addressee) is the *marked* situation. Gaze aversion is typical for speaking. Looking at the addressee while speaking can thus serve a signalling function. This could explain why SpeakerGaze affects lexical alignment. The other way around, AddresseeGaze (i.e. addressee fixating speaker) does not generate this effect. This might be explained because AddreseeGaze is

default for addressees. Addressees look at their speaking partner (nearly) all the time anyway. AddresseeGaze thus does not have the marked, signalling function SpeakerGaze has. Interestingly, the SpeakerGaze effect does not appear to transfer to the target. In related studies (Postma et al. 2013, Wang et al. 2011, 2014) prime and target immediately follow each other and the gaze behaviour during the two is kept constant. In our data this is different: prime and target can (and also do) differ in terms of gaze behaviour. Participants' fixations during prime and target are independent, and the SpeakerGaze effect only resides in the prime. This again hints at the relevance of the signalling function of gaze. Only during the prime can SpeakerGaze have a signalling function. When the participant produces the target word, it is too late for the speaker in the prime (i.e. the addressee in the target) to still use gaze to highlight anything in his initial message. In his role as addressee he can only signal attention to what the speaker in the target is saying. Furthermore, the audience effect that gaze might constitute, does not seem to play an important role in our data. If the social aspect of 'being looked at' would affect lexical alignment, we would expect this to typically (or at least also) occur when producing the target word. However, AddresseeGaze during the target does not affect lexical alignment.

What this case study demonstrates is how gaze enhances alignment. Although there is a lot of idiosyncratic variation in gaze behaviour (Cummins 2012, Kendon 1967) and although gaze serves many different functions at the same time (even ranging between mere perception and communicative signalling), we measure an effect of SpeakerGaze on lexical alignment. In our set-up of uncontrolled, spontaneous speech we cannot discriminate between when exactly gaze serves cognitive, social or communicative functions, but we do find evidence that communicative functions of eye gaze, viz. its signalling function, and not cognitive or social functions alone, are good predictors for lexical alignment. What experimentally controlled studies such as Wang et al. (2011, 2014) or Postma et al. (2013) demonstrate, is an immediate and socially or cognitively motivated impact of eye gaze on alignment. This case study reveals there is a mediated and communicatively motived impact as well.

3.1.2 Gaze and gestural alignment

Above, we have demonstrated a link between gaze behaviour and lexical alignment. In this case study we will test whether the same link holds true for gaze and gestural alignment. Compared to the previous section, we will roughly use the same set-up and analysis but the crucial difference lies in the type of gaze behaviour we (are able to) measure. For speech, the only relevant gaze behaviour to be measured was "fixating the face" or "not fixating the face". We will do this for gesture as well, and hypothesise that if a speaker looks at the addressee while performing the gesture, the addressee will be more likely to align to that gesture, compared to when the speaker averts his gaze during the gesture production. Because the articulators of gesture, i.e. the arms and hands, are more visible than the articulators for speech, it might also be relevant to measure whether or not speaker and/or addressee fixate those gestural articulators. In other words, we will also study the link between fixation on gestures and gestural alignment, and hypothesise that if speakers or addressees have looked at the gesture in the prime, addressees will be more likely to perform an aligned gesture in the target.

This research topic ties in with studies by Wang and colleagues (2011, 2014). These researchers found that when addressees are being looked at by a speaker in a video, they are faster at copying the gesture the speaker just performed. However, no such effect was found when speakers fixated their own gesture, i.e. addressees were not faster in copying the gesture if that gesture was fixated by the speaker, compared to when it was not. Perhaps needless to repeat, but we again want to stress that our study crucially differs from Wang et al. (2011, 2014) in two respects. First, our dependent variable is gestural alignment. We study whether or not two gestures in a prime-target pair are the same, whereas Wang and colleagues study the reaction time to a stimulus gesture. Second, the interactional setting is very different. We study gestural alignment in face-to-face conversations, whereas Wang and colleagues study gestural reaction times in non-interactional experimental tasks.

Next to following up on Wang et al. (2011, 2014), the case study in this section also ties in with the work by Gullberg & Kita (2009). In that study participants were asked to watch videos of people telling a cartoon

story. The people in the video were talking to a live addressee who was not visible on the video images. During this story telling spontaneous co-speech gestures occurred. The researchers were interested in a subset of target gestures in the videos: gestures encoding spatial events with the speaker either focussing or not focussing on his/her own gestures, and with the spatial information present in the speech or not present in the speech of the story teller. The participants watching the videos were eye-tracked and asked to draw a selection of target scenes in the cartoons after having watched all of the videos. This was done to dig into the *information uptake* of the target gestures. The crucial question here was whether in the drawing task, fixated gestures (either by the speaker or the addressee) were more adequately drawn, in terms of spatial dynamics, than non-fixated gestures. The main findings, relevant to the present work, in Gullberg & Kita (2009) can be summarised as follows:

(i) Addressees do not focus on many of the speakers' gestures.
(ii) If addressees do look at the speaker's gesture, then often that speaker has focussed on his own gesture (i.e. gaze cueing).
(iii) If a speaker has looked at his own gesture, the addressee will retain more of the information encoded in that gesture.
(iv) If an addressee has focussed on the gesture of the speaker, that addressee will not retain more of the information encoded in that gesture.

Drawing on (i) and (ii), the results above indicate that eye gaze has a strong cueing effect: only if speakers focus on their own gestures, addressees focus at those gestures as well. What (iii) and (iv) demonstrate, is a relation between gaze behaviour and information uptake. If speakers fixate their own gestures, and not if addressees focus on those gestures, the information uptake is higher, i.e. the participants retain more of the spatial information encoded in the gesture. What Gullberg & Kita (2009) do not study, and where this case study fits in, is whether fixations on gestures (either by speaker or by addressee) affect subsequent gesture production. The research questions, then, for this case study can be schematised as follows:

(i) Do interlocutors focus on a minority of their partner's gestures also in face-to-face dialogues?
(ii) Is gaze cueing crucial also in face-to-face dialogues?
(iii) If speakers or addressees focus on gestures during the prime, are addressees then more likely to align to that gesture in their subsequent gesture production?
(iv) If speakers or addressees focus on each other's face (either during the prime or the target), are addressees then more likely to gesturally align?
(v) As was the case for lexical alignment, is gestural alignment enhanced by speakers or addressees fixating the face of their conversational partner?

METHOD & ANALYSIS

Just as in the previous case study on the relation between gaze and lexical alignment (3.1.1), we use the data from the animation description task in the Insight Interaction Corpus (see Chapter 2). Since the present study is concerned with the coupling of gesture and gaze, it is important to note we single out one specific type of gesture, viz. *depictive gestures*. These are used by the participants to represent the target objects that are present in the animation videos. All other gestures, like emblems or beat gestures, are not part of the data set for the present analysis. In testing our hypotheses, we compare two factors: the alignment between adjacent representational gestures, and the eye gaze of the interlocutors.

To quantify gaze behaviour, we use the gaze annotation from the Insight Interaction Corpus (see 2.3.5) to determine whether or not speakers or addressees are focussing on the target gestures or on each other's faces. The annotation code GEST indicates addressees are fixating the gesture performed by the speaker, OWN indicates speakers fixating their own gestures, and FACE indicates participants fixating their partner's face. Especially in considering fixations on gestures, we need to point out that any study on visual fixations (with or without the help of eye-tracking tools) can only provide positive evidence: if there is a fixation on a given object, we can assume the participant has cognitively processed the visual stimulus. However, if there is no fixation, it cannot be ruled out that the

participant still has processed the stimulus. This is due to the human peripheral vision, which allows information uptake without explicit fixations within an angle of 120° (Duchowski 2007: 29-32). For example, as is clear from eye-tracking research in sign language (Muir & Richardson 2005), signers hardly ever fixate their interlocutors' hands, while they obviously do 'see' what their conversational partners are expressing with those hands. Because peripheral vision allows perception without fixation, we should take care in how we interpret the data in this case study, but also beyond.

To measure gestural alignment we want to answer the question 'are prime and target gesture aligned?'. This is problematic given the multidimensionality of gesture. For example, if two gestures have the same hand shape and finger orientation but a different palm orientation, can they be considered as fully aligned? In their work on gestural alignment Bergmann & Kopp (2012) acknowledge this multidimensionality and calculate gestural alignment on one of five separate gesture features (representation technique, handedness, hand shape, palm orientation, finger orientation and wrist movement). For this study we only use one of those features, viz. representation technique. As explained in Chapter 2, for the annotation of that feature we adopt the typology of depictive gestures by Streeck (2008: 292-295), who distinguishes gestural depiction methods such as modelling (hand as a token for an object), bounding (hands indicate sides or edges of an object), drawing (fingers draw lines that represent the outline or path of an object), handling (hands enact a prototypical usage of the represented object), etc. By only using representation technique as a basis for our dependent variable of gestural alignment, we miss out on some formal features such as hand shape or palm orientation. However, since we are crucially interested in the question 'are prime and target gesture *the same*?', adopting a holistic approach and only considering representation technique appears to be justified. In this study we do not want to measure whether gaze behaviour affects palm orientation, but whether it affects gesture production as a whole. If prime and target gesture differ in representation technique, they also differ in many formal features. Which formal features exactly is not the topic of this study.

Measuring gestural alignment is very analogous to our measuring lexical alignment (see 3.1.1). To illustrate how we compare prime and

target gestures, consider the example below. In this example we show how we measure gestural alignment for the target object DOOR. The verbal references to that target object (marked in bold and red) are accompanied by gestural references (marked with red circles in Fig. 16).

S2 There's a **door**.
S1 A black **door**.
S2 Yes.
S1 Yes, well, a **hole**.
S2 And there's a guy standing in front of the **hole** with his hands in his pockets.

Fig. 16: The target object DOOR is represented four times in this example

In the example above, there are two prime-target pairs (see green rectangles in Fig. 17 below). Those pairs are defined as adjacent gestures produced by different speakers. For example, the second and third gesture in the example (see Fig. 17) are adjacent, but they are produced by the same speaker. Similarly, gesture one and three are no prime-target pair: although they are produced by different speakers, they are not adjacent. As mentioned above, in order to label two gestures in a prime-target pair as aligned ones, for this study, we only consider the representation technique (according to Streeck 2008). This means that for the first interactional pair in the example, we measure alignment in the representation technique

'drawing', although the two gestures are not identical (the most prominent difference being that the girl uses two hands and the boy only one hand). For the second interactional pair, we see a parallel issue: the finger orientation and tension in the hand shape differ between the two speakers, but we still consider it to be an instance of gestural alignment because the representation technique is identical (i.c. modelling).

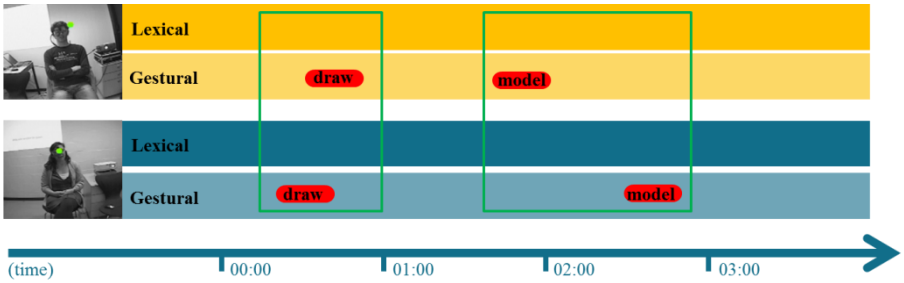

Fig. 17: Gestural references to the target object DOOR *in which both prime-target pairs are aligned*

We identified all prime-target pairs (n=536) for all of the speakers and coded them for the two factors discussed above: gestural alignment and gaze behaviour. In some pairs a gaze shift occurs during the production of a prime or target gesture. Because we only want to include pairs in which there either is a full fixation or a full gaze aversion during the entire duration of the gesture, we omitted those gaze shift cases from the data set. From the initial 536 prime-target pairs, we thus keep 417. In the sample conversation above (see Fig. 16 and 17), both prime-target pairs are aligned, there is no fixation on own (SpeakerOwnGest) or other (AddresseeGest) gestures during the prime, and participants focus on each other's face (SpeakerGaze & AddresseeGaze) both during the prime and during the target. Our scoring table for this brief example, then, would look like this:

	Alignment	Speak Gaze Prime	Speak Gaze Target	Address Gaze Prime	Address Gaze Target	Speak Own Gest	Address Gest
pair1	1	1	1	1	1	0	0
pair2	1	1	1	1	1	0	0
…	…	…	…	…	…	…	…

Table 2: Example coding scheme of the dependent factor alignment and the independent gaze factors

Parallel to the previous section on lexical alignment (see Table 1), we also calculated from the gaze data whether or not there was *eye contact* during prime and during target, and added that as a factor. We then computed mixed effects models to uncover whether the gaze behaviour of the speaker (focussing on his own gestures or on the addressee's face) or the gaze behaviour of the addressee (focussing on the speaker's gestures or face) are good predictors of whether or not prime and target gestures are aligned. To account for variation in alignment rates that is due to specific dyads or objects we treat *dyads* and *objects* as random factors in our models. To account for collinearity issues, we tested every possible interaction between our independent variables and found that none of those interactions exceeded a Cramer's V value of 0.14.

RESULTS

In answering the first research question (i), we found that also in face-to-face dialogues very few gestures get fixated. On a total of 1770 depictive gestures only 3.7% are fixated by the addressee (AddresseeGest), and 4.0% by the speaker (SpeakerOwnGest). Gullberg & Holmqvist (2006) found 7.4% of AddresseeGest, but because they do not report absolute frequencies and because we restricted ourselves to one subtype of gestures (viz. depictive gestures), we cannot statistically compare the obtained results.

Research question (ii) was on the relation between AddresseeGest and SpeakerOwnGest, viz. the so-called *gaze cueing* effect. Our data provide evidence of a strong gaze cueing effect: 46.5% of the SpeakerOwnGest fixations are followed by an AddresseeGest fixation. This

means that nearly half of the time the gaze cueing is successful[22]. This result appears to be significant: a mixed effects model with AddresseeGest as dependent factor, SpeakerOwnGest as fixed factor and *dyad* as random factor reveals that the gaze behaviour of the speaker is a good predictor for that of the addressee ($z=4.85$, $p<0.001$).

Research question (iii) addressed the relation between gaze behaviour of the speaker and gesture behaviour of the addressee. Our hypothesis that addressees align more to the gestures of the speakers if those speakers focussed on their own gestures was not confirmed. Fig. 18 shows there is hardly a difference in alignment scores for the prime-target pairs in which the speaker looked at his own gesture in the prime (SpeakerOwnGest+), compared to when the speaker did not look at his own gesture (SpeakerOwnGest-).

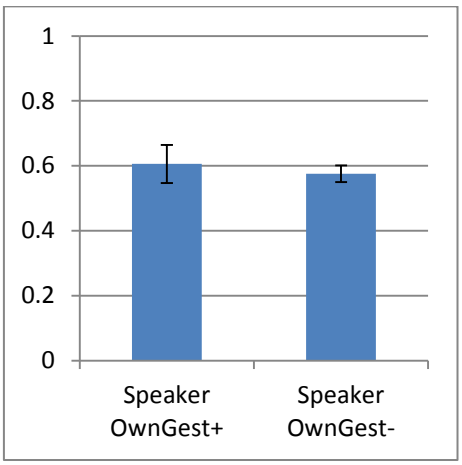

Fig. 18: Average alignment rates for speakers fixating their own gestures (SpeakerOwnGest+) and speakers not fixating their own gestures (SpeakerOwnGest-). Error bars indicate standard error.

[22] For us, *successful* gaze cueing is a mere technical matter of co-occurring speaker and addressee fixations. We do not take into account which intentions speakers might have when fixating their own gestures. Those intentions might well be to explicitly invite the conversational partner to look at the produced gesture, but speakers can have many other reasons to focus on their own gestures.

Addressing research question (iv) we do find a significant difference in gestural alignment rates. If addressees have fixated the prime gesture (AddresseeGest+), they align to that gesture in 73.9% of the cases, whereas if there is no such fixation (AddresseeGest-) the average alignment rate is only 55.3% (see Fig. 19).

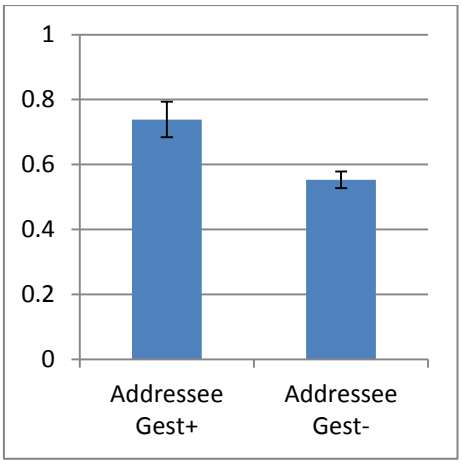

Fig. 19: Average alignment rates for addressees fixating the prime gestures (AddresseeGest+) and addressees not fixating the prime gestures (AddresseeGest-). Error bars indicate standard error.

Zooming in further on research questions (iii) and (iv), what has not yet been addressed is the relation between SpeakerOwnGest and AddresseeGest. Comparing those two factors creates four possible gaze configurations for which we calculated the average gestural alignment scores. Fig. 20 visualises those four configurations: (a) speaker and addressee both fixating the prime gesture, (b) only speaker fixating the prime, (c) only addressee fixating the prime, (d) speaker nor addressee fixating the prime. Fig. 21 shows the average alignment scores for those four configurations.

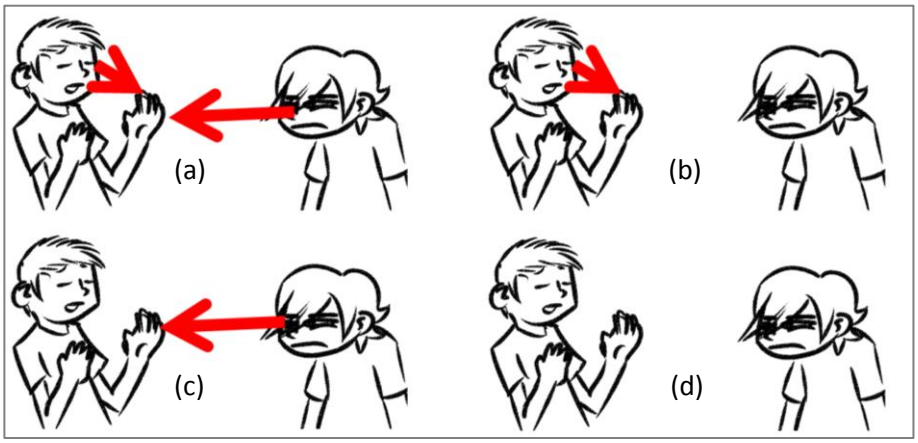

Fig. 20: Four possible gaze configurations when combining the two factors SpeakerGaze and AddresseeGaze

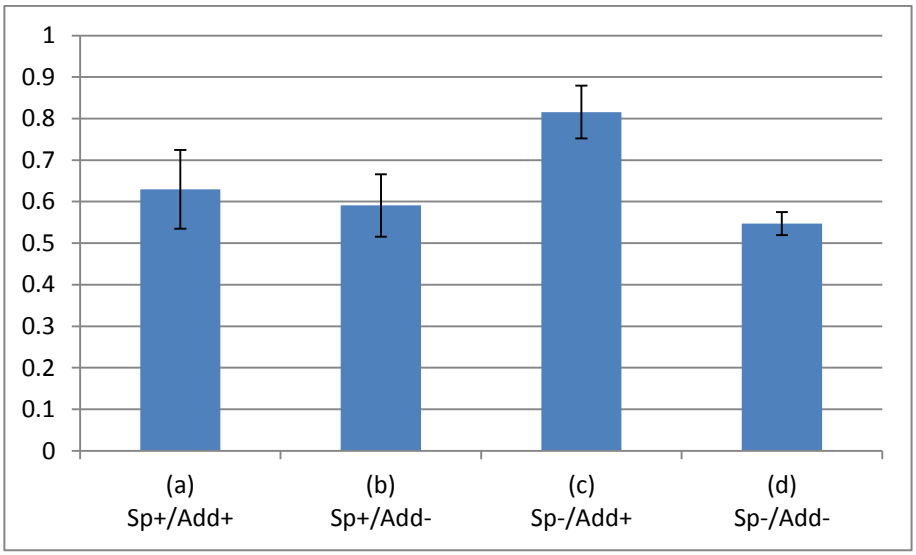

Figure 21: Average scores for gestural alignment across two factors: SpeakerOwnGest (speaker has or has not focussed on his own hand gesture) and AddresseeGest (addressee has or has not focussed on the speaker's hand gesture). Error bars indicate standard error.

As was already apparent from the previous results, the average scores in the AddresseeGest+ cases ((a) and (c)) are higher than their AddresseeGest-

counterparts ((b) and (d)). This means that if addressees have focussed on the speaker's hand gesture, they produce more aligned gestures than if they have not focussed on the speaker's gestures. What stands out here is the SpeakerOwnGest+/AddresseeGest+ configuration (a). If both SpeakerOwnGest+ (Fig. 18) and AddresseeGest+ (Fig. 19) correlate with higher scores on gestural alignment, then we would have expected the SpeakerOwnGest+/AddressseeGest+ configuration (a) to show the highest alignment scores. Figure 21 shows this is not the case.

To test the significance of the effect of SpeakerOwnGest, AddresseeGest, and the interaction between them, we computed a mixed effects model where gestural alignment was the dependent factor, SpeakerOwnGest and AddresseeOwnGest the independent factors, and DYAD and OBJECT the random factors. This revealed a significant main effect of AddresseeGest ($z=2.664$, $p=0.007$), qualified by an interaction with SpeakerOwnGest ($z=-1.914$, $p=0.05$). There was an interaction between the factors because only in the SpeakerOwnGest- cases the difference between AddresseeGest+ and AddresseeGest- was significant. In the SpeakerOwnGest+ cases, there was no such difference. In other words, only when speakers do not fixate their own gestures, it matters (in terms of gestural alignment scores) whether the addressee looks at his partner's gestures. If the speaker does fixate his own gestures, the gaze behaviour of the interlocutor no longer correlates with significantly higher gestural alignment scores.

Unlike what we found for lexical alignment, and answering research question (v), we found no effect of SpeakerGaze or AddresseeGaze (i.e. fixating the partner's face by resp. the speaker and the addressee) on gestural alignment. As is already clear from Fig. 22, SpeakerGaze nor AddresseeGaze, either during the prime or during the target, enhances gestural alignment. A mixed effects model with gestural alignment as dependent variable, SpeakerGaze and AddresseeGaze as independent factors and DYAD and OBJECT as random factors confirmed that none of the independent factors reached significance. Similarly, a mixed effects model with the same dependent and random factors, but eye contact during prime and during target as independent factors, failed to reach significance

as well. Together these tests reveal that fixations on the face, regardless of when and by whom, are irrelevant in explaining gestural alignment.

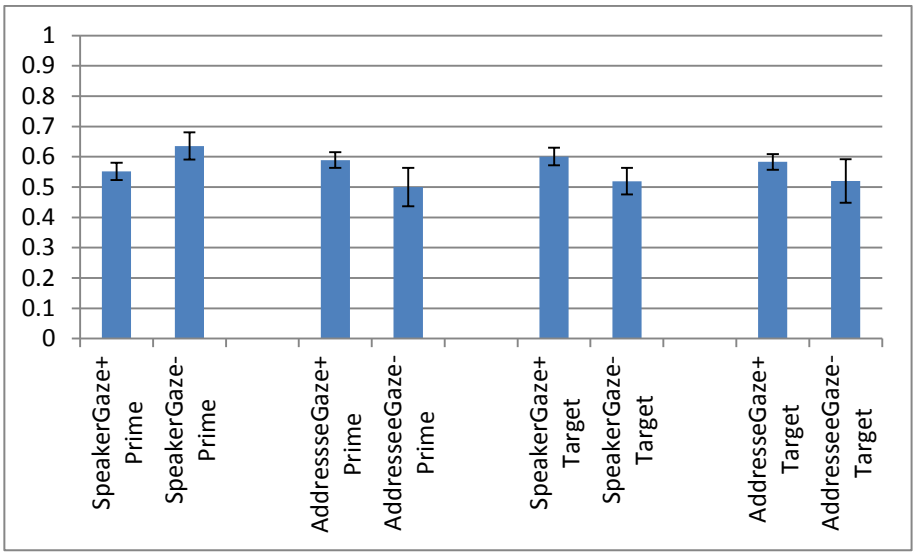

Fig. 22: Average scores for gestural alignment for SpeakerGaze and AddresseeGaze during prime and target. Error bars indicate standard error.

DISCUSSION

Fixations on gesture do not occur often. Only in roughly four percent of the cases or less than one percent of the time participants in our corpus fixated a gesture. However, if it does occur, something happens. First, SpeakerOwnGest and AddresseeGest co-occur often: if a speaker looks at his own gesture, nearly half of the time the addressee does the same thing. This gaze cueing effect was already pointed out (Emery 2000, Frischen et al. 2007, Lachat et al. 2012), but to the best of our knowledge this is the first account of the phenomenon in which eye-tracking is used in a face-to-face interactional setting. We thus provide evidence that the observations made in strictly controlled lab settings can be stretched to spontaneous speech.

The second relevant finding of this study pertains to the multimodal relation between fixations on gestures and subsequent gestural behaviour. Regarding this relation, the main results can be summarised as follows: SpeakerOwnGest alone (i.e. the speaker fixating his own gesture) does not affect gestural alignment, AddresseeGest (i.e. the addressee fixating the

speaker's gesture) does significantly co-occur with higher scores for gestural alignment, but it only does so in the SpeakerOwnGest- cases (i.e. when the speaker does not fixate his own gesture). That we did not find an effect for SpeakerOwnGest ties in with what Wang et al. (2014) found in their reaction time experiment: addressees looking at actors in a video were not faster in copying target gestures if those gestures were fixated by the actor, compared to when they were not fixated. However, our results are quite different from Gullberg and Kita's (2009) findings. They showed how SpeakerOwnGest did and AddresseeGest did not have an effect on gestural information uptake. In our study, it appears to be the other way around. Combining the results of both studies: if a speaker fixates his own gesture, addressees retain more of the information in that gesture, but they are not more likely to align in subsequent gesture production. The other way around, addressees that fixate a prime gesture do not retain more of the information, but they are more likely to gesturally align to the speaker. This might mean that gesture information uptake and gestural alignment are too disparate phenomena to be compared. In other words, maybe a higher information uptake does not lead to more gestural alignment. This assumption, however, is in contraction with the foundations of priming: if fixating a gesture as an addressee leads to more information uptake, then that should lead to a higher likelihood of gestural alignment (cf. the notion of *activation level* in Pickering & Garrod 2004, 2006). Therefore, the difference in results between Gullberg & Kita (2009) and this study might be relevant, and can be explained by the difference in conversational setting: watching a speaker telling a story in a video in the former study, and a face-to-face collaborative task in the latter. In this vein, the divergent results could be indicative of the fact that identical gaze events (SpeakerOwnGest and AddresseeGest) serve different functions in different conversational settings. This interpretation is of course speculative and future research is needed to substantiate this hypothesis

What then can we conclude from the fixations on gesture in our data set? First, although gaze appears to be an efficient tool for a speaker to make his addressee focus on a gesture he is performing (cf. *gaze cueing*), SpeakerOwnGest is not functional in terms of gestural alignment. Given that fixations on gesture by addressees and not by speakers correlate with

higher gestural alignment scores, it may not be so surprising that gaze configuration (a), viz. both speaker and addressee are fixating the gesture, does not correlate with the highest alignment score and that gaze configuration (c), viz. only the addressee fixating the gesture, does. To make a somewhat simplifying comparison: given that people liking apples get high grades and people liking pears do not get high grades, we would expect people only liking apples (cf. gaze configuration (c)) to get higher grades than people liking both apples and pears (cf. gaze configuration (a)). Comparing the results between gaze configurations (a) and (c), we could hypothesise that addressees may have different reasons for fixating a gesture. One reason would be fixating a gesture because the speaker is fixating the gesture. Here the addressee is being coerced, the gesture fixation is externally triggered. Another reason could be to better process (the physical structure of) the gesture. In this vein the gesture fixation is internally motivated. Perhaps, more than an external trigger, an internal trigger to gesture fixation makes gestural alignment more likely. Of course, whatever the reasons are why addressees fixate gestures, we can only speculate on those reasons in this study due to lack of experimental control. Notwithstanding this speculation, we have at the very least demonstrated that gesture fixation is relevant in explaining gestural alignment in face-to-face conversations. Even amidst the many different functions gaze has, functions that are not evoked or used during experimental, non-dialogic tasks, we measured a significant effect[23] of the subtle interplay between SpeakerOwnGest and AddresseeGest on gestural alignment.

A second main issue we addressed in this case study pertains to the relation between gestural alignment and gaze fixations on the face. For none of the gaze factors we coded, i.e. SpeakerGaze and AddresseeGaze

[23] Although the experimental set-up does not allow us to measure a *causal* link between gaze behaviour and gesture behaviour, we do observe a temporal contingency between the two: for each instance in our data set the gaze behaviour (viz. focus on gesture) precedes the gestural behaviour (viz. target gesture). We cannot claim that the gestural alignment we measure happens *because of* the gaze behaviour, but we do want to stress that at least there is a clear temporal contigency between the two.

during both prime and target, we found a correlation with higher gestural alignment rates. This is in contrast with what Wang et al. (2011, 2014) found in their reaction time experiments: they found evidence that when SpeakerGaze during the prime occurred, addressees were significantly faster at copying the gesture they just perceived. A factor contributing to the conflicting results could be the effect of *distance*. In Wang et al. (2011, 2014) prime and target gesture always immediately follow each other. Due to the natural conversation in our study, there is a variable, but crucially a much larger distance between prime and target. The effect of SpeakerGaze during the prime might therefore wear off in many of our prime-target pairs. The absence of an effect for SpeakerGaze is not only in contrast with the results obtained by Wang and colleagues, but also with the previous case study (see 3.1.1) on lexical alignment. Apparently, the gaze behaviour of the speaker during the prime does not matter in terms of gestural alignment rates, but it does in terms of lexical alignment rates.

Combining the results of this case study and the one in 3.1.1 on lexical alignment we see some interesting differences. Whereas lexical alignment is enhanced by gaze behaviour of the speaker, gestural alignment is enhanced by gaze behaviour of the addressee. In the introduction to this case study we noted that gaze can serve both perception and production purposes. We use it to perceive the world around us and to convey meaning during interaction. This dichotomy appears to be relevant in explaining the relation between gaze and multimodal alignment. The production aspect of gaze, viz. the signalling or highlighting function of it, appears to be crucial in explaining lexical alignment. Higher lexical alignment rates correlate with the speaker looking at the addressee during the prime. The perception aspect of gaze, viz. the (cognitive) focus of attention, appears to be linked with gestural alignment. Higher gestural alignment rates correlate with the addressee fixating the speaker's gesture during the prime.

The differences in results between Gullberg & Kita (2009) and Wang et al. (2011, 2014), and this case study are indicative of the many functions of eye gaze in communication and of the intricate relationship between gaze behaviour and alignment. Interlocutors in conversations can fixate their partners or their (partners') gestures for many different reasons:

disambiguating, gaze cueing, signalling uncertainty, deictic referencing, etc. Using mobile eye-tracking allows us to measure visual fixations in great detail, but of course without getting direct access to these different conversational functions. Also, so far, we ignored a whole range of parameters that may influence addressees' (overt) attention to an interlocutor's gesture or the likelihood of establishing aligned representations at the non-verbal level. McNeill (2006) and Gullberg & Kita (2009) provide an overview of such factors including social status, interpersonal stance, speaker information structure, shared common ground and the physical properties of the gesture. Apart from those factors, also the time difference between the fixation onset and gesture onset, the fixation duration, co-occurring verbal cues, the number of preceding gestures that were or were not fixated, etc. might be parameters with explanatory potential as well. In the next section we will try to include some of those factors in an attempt to provide a more accurate, and more interactionally grounded account of the multimodal alignment that remains our dependent variable.

3.2 Case study 2: a multifactorial account of lexical and gestural alignment

In the previous section (3.1) we showed how gaze can affect lexical and gestural alignment. However, other factors may predict whether or not people in interaction align as well. After all, as already pointed out by Brennan & Clark (1996), *historical* factors, viz. contextual factors, are necessary to fully account for alignment. Ahistorical factors alone, such as the interplay between informativeness and conciseness in lexical choice, are insufficient to adequately describe the phenomenon (see section 1.3.1 for a detailed discussion on this). In this second case study we will show how the conversational context shapes alignment. The case study is multimodal in two respects. First, we check whether the same factors predict alignment for different modalities (i.e. speech and gesture). Second, we check whether behaviour at one modality is a good predictor for alignment at another.

3.2.1 Introduction & research questions

Alignment as a driving force in interaction is not restricted to the simple repetition of lexical items or syntactic structures in adjacent turns in conversation. Rather, it is a dynamic contextually embedded phenomenon, in which different semiotic channels, including gesture, posture and gaze, are tightly coordinated between the interlocutors. With this case study we want to address two interrelated questions that have not received substantial attention in the literature. First, which factors may explain the occurrence of interactive alignment (sequences) in longer stretches of face-to-face interaction? On the basis of different (psycholinguistic) models of dialogue, we select a series of variables pertaining to the (social) dynamics of the interaction (including speaker dominance, the temporal distance between utterances and cumulative priming) and try to model their relative impact. Second, does a similar pattern of interactive alignment emerge across different modes of representation? In other words, do the same factors predict gestural and lexical alignment?

In trying to further pinpoint our research questions, we start from Brennan & Clark (1996), who argued that lexical entrainment is subject to the frequency-of-use hypothesis. They state that “two partners should rely

more on a conceptualisation precedent the more firmly it has been established" (ibid.: 1498). This prediction was confirmed in their study, based on a picture-naming task. What the study did not address, however, is the question whether:

(i) this effect of *cumulative* or frequent use pertains to a co-participant's or speaker's own linguistic choices in the preceding trials. In other words, does it matter who produced the precedents in the interaction and how often?
(ii) other factors than frequency of use may help to predict lexical or gestural alignment/entrainment.

In addressing non-verbal alignment, Bergmann & Kopp (2012) and Louwerse et al. (2012) start from a behavioural approach towards the topic of alignment. They measure the occurrence of gestural alignment and synchronisation across speakers, independently of the conceptual representation linked to that gesture. In other words, other than in lexical studies such as Brennan & Clark (1996), they focus solely on a comparison of the physical form of adjacent gestures (e.g. hand shape, orientation etc.) and ignore the question whether these adjacent gestures (help to) express the same concept (e.g. two subsequent gestures depicting the same object). In this case study we want to explore the questions whether:

(iii) taking a representational rather than a purely form-based approach to gestural alignment generates the same results;
(iv) the occurrence of aligned gestural depictions across speakers is driven by the same explanatory factors as in lexical alignment (e.g. cumulative priming, question (i));
(v) lexical alignment typically coincides with gestural alignment (i.e. does lexical alignment correlate with gestural alignment or the other way around?).

3.2.2 Method & analysis

For this case study we use the same data set as for the previous studies (3.1), i.e. the animation description task in the Insight Interaction Corpus

(see Chapter 2). Also, our annotation of the dependent variables, lexical and gestural alignment, is analogous to the case studies in the previous section: we only consider prime-target pairs and score whether or not interlocutors use the same word/gesture in prime and target. Because it is crucial to the interpretation of our results, we repeat that, unlike Bergmann & Kopp (2012) or Louwerse et al. (2012), our prime-target pairs always have the same referent. In doing so, we want to tap into alignment at the referential level ("do interlocutors use the same words/gestures to refer to the same things?") rather than at the purely behavioural level ("do interlocutors use the same lexical/gestural formal features regardless of what they are referring to?"). Before turning to the description of our independent variables, we want to address how we tackled two problematic issues that are relevant in interpreting the results: content confound and baseline comparisons.

OVERCOMING THE CONTENT CONFOUND PROBLEM

As pointed out above, we performed a digital coding for the lexical items and gestures (plus or minus alignment). In some communicative settings however, alignment seems to be almost unavoidable. Du Bois (2010: 31) refers to this issue as the *content confound* and raises the methodological question that alignment "may have simply been imposed upon the speakers by factors not entirely under their control, such as the current topic (the subject matter under discussion) and the limited set of words that the language provides for expressing this content. When two speakers engaged in conversation use the same words, is not that just because they're talking about the same topic?".

It is plausible that if a given language only offers one lexical option to label a certain object, it is impossible for them not to align in naming that object (except for the case of circumscriptions, Costa et al. 2008). Vorwerg (2013: 152) makes the same claim, but states it the other way around: "the existence of a variety of linguistic means to express a particular idea or message both allows for and necessitates verbal attunement in communicative interaction." Du Bois (2010) uses the example of 'liver' as a referent that has no common lexical alternatives, so that if interlocutors are talking about this topic they have no option to not lexically align. However,

if two speakers in two consecutive turns use the lexical item 'convertible' to refer to the vehicle their friend owns, both speakers had at their disposal a vast repository of possible lexical labels to name that vehicle (vehicle, car, Chevrolet, sportswagon, and so on) and thus had multiple options to not align. Since alignment is our dependent variable (and we are counting either plus or minus alignment cases), we wanted to rule out as much as possible the cases where the content confound makes it impossible for interlocutors not to align.

Prior to recording the Insight Interaction Corpus a pre-test was performed, in which all of the target objects in the video animations were checked for sufficient onomasiological variation potential (high lexical choice variability, cf. Brennan & Clark 1996). In a labelling game, students were asked to name a set of objects they were shown. Only if there was sufficient variation and spread of lexical labels per object, that object was selected for the video animations. That this labelling game yielded satisfactory results, will also be clear from the results section.

BASELINE COMPARISON

As explained in the previous section, we maximally tried to avoid the content confound issue in our study. To further rule out that that we are measuring co-incidental co-occurrences of lexical items and gestures, we created a baseline comparison for our results and calculated whether there was a significant difference between that baseline and the actual results.

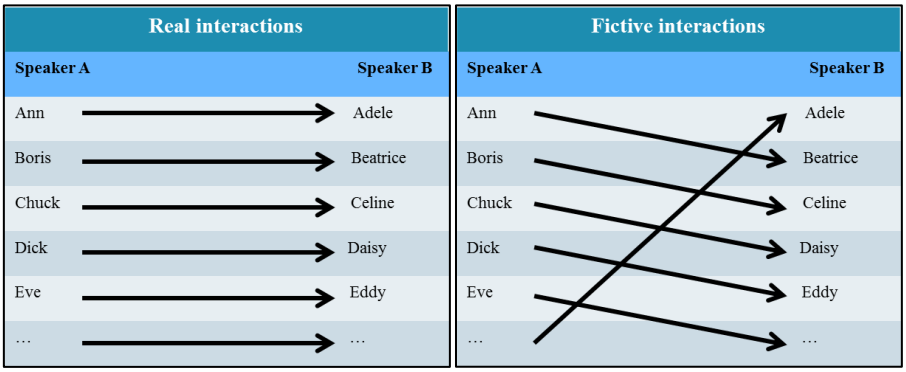

Fig. 23: The couples in the real interactions are decoupled and shuffled in the fictive interactions

The baseline in our study is a set of fictive interactions[24]. We obtained these fictive interactions by shuffling speakers so that we matched the time-aligned annotation strings of speaker A in pair 1 with that of speaker B in pair 2, speaker A in pair 2 with speaker B in pair 3, and so on (see Fig. 23). Because the interlocutors in those fictive interactions are still referring to the exact same target objects, it was possible to apply the same measuring techniques for lexical and gestural alignment. The results of measuring alignment in those fictive, shuffled interactions will form a baseline for the results of the actual conversations. To increase the reliability of this baseline we would ideally create the maximum amount of fictive interactions, i.e. to connect each speaker A with each of the speakers B from the remaining 14 interactions. However, for the scope of this case study, and because the annotation process is not an automatic one, we randomised the conversational partners four times, creating 60 (fifteen dyads in the corpus that got shuffled four times) fictive interactions. This is comparable to what Bergmann & Kopp (2012) and Howes et al. (2010) did, who shuffled conversational partners just once, creating one fictive dialogue per real dialogue.

INDEPENDENT FACTORS

So far, we have only discussed our method for measuring the dependent variable. Now we turn to the independent variables that might be good predictors for interactive alignment to occur. Each of these predictors can be linked to a specific hypothesis (for a schematic overview, see Table3). It is important to note that all of the factors presented here will be used in predicting the alignment score at both the lexical and the gestural level.

[24] See Richardson & Dale (2005), Howes et al. (2010) or Bergmann & Kopp (2012) for a comparable baseline condition creation in conversational data.

Code	Research question *(hypothesis)*
distance	Are words/gestures closer to each other more aligned? *(y)*
position	Is there more lexical/gestural alignment towards the end of the conversations? *(y)*
block	Is there more alignment towards the end of the experiment? *(y)*
prime-self	Will speakers align more if they already used the same word/gesture themselves? *(n)*
prime-other	Will speakers align more if their interlocutor already used the same word/gesture? *(y)*
words	Do the most talkative speakers align the most? *(n)*
1_mention	Do the topic introducing speakers align the most? *(n)*
gaze	Does gaze affect alignment? *(y)*

Table 3: Overview of the independent variables

Temporal distance and position

The factor *distance* has already been shown to play a role in gestural alignment in an interactional setting of instruction-giving (Bergmann & Kopp 2012). For this study we calculated temporal *distance* as the time difference between (the offset of) the prime and (the onset of) the target of a prime-target pair. The hypothesis is in line with the results in Bergmann & Kopp (2012): if prime and target are closer together, they have a higher chance of being aligned.

A second factor relating to the temporal dynamics of alignment is temporal *position*: for each interactional pair we calculated the relative position in the conversation[25]. Note that this is linked to the factor *block* but the two factors should be treated separately: whereas *block* expresses the

[25] A *conversation* refers to one of the fifteen discussions that happened after the participants saw a video animation. In a fictive example of a 280 second conversation, if the target part of an interactional prime-target pair occurred at second 140, this would be exactly halfway into the conversation, so the (relative) value for the temporal position in this case would be 0.5.

position within the entire experiment[26], temporal *position* expresses the position within each conversation. Our hypothesis corresponds to the findings of Louwerse et al. (2012: 15): "the more interlocutors interacted with each other, the more they synchronised matching behaviors with one another"; i.e. the further into a conversation (and into the experiment), the more alignment we expect.

Cumulative priming

A second group of factors we included into our model concerns the effect of cumulative or structural priming. This effect has already been demonstrated for, for example, phonetic alignment (Babel 2009, Lewandowski 2012), lexical alignment (Brennan & Clark 1996) and syntax alignment (Bock & Griffin 2000, Pickering & Ferreira 2008). The hypothesis is that the more the interlocutors hear/see a word/gesture, the more likely it will be that they align to that word/gesture. What the studies referred to all measure is whether a repetition of a stimulus affects alignment rates. In this case study, which is concerned with alignment in face-to-face conversation, we want to be able to differentiate between cumulative priming by the addressee (*prime-other*) and cumulative priming by the speaker (*prime-self*). To make this set of parameters sufficiently clear, consider the last interactional pair (rightmost green rectangle) in the example given in Fig. 13.

- *prime-self*: how many times, before the interactional prime-target pair, has the current speaker used the same word/gesture? (once: prior to the interactional pair the girl referred to cat with "pussy" 1 time)
- *prime-other*: how many times, before the interactional prime-target pair, has the other speaker used the same word/gesture the current speaker is using? (once: prior to the interactional pair the boy referred to cat with "pussy" 1 time)

[26] In the data set the 15 conversations were split in three *blocks* of five conversations, to check the calibration of the eye-tracker. The factor *block* thus has three values: block 1, 2 and 3.

With the set of factors above we not only measure whether or not there is a cumulative effect of priming, we also measure whether this possible cumulative effect is speaker-tied (*prime-self*) or addressee-tied (*prime-other*). It is important to note here that we are not measuring self-alignment. Our dependent variable throughout the entire dissertation remains the same: interactional alignment. We only look at interactional prime-target pairs in which the speaker in the prime is different from the speaker in the target. What we do factor in here, is how self-priming affects interactional alignment. This is crucially different from measuring how self-alignment relates to interactional alignment.

Dominance

Social and emotional factors have been shown to determine the occurrence and rate of alignment phenomena (a.o. Chartrand & Bargh 1999, Hove & Risen 2009, Lakin et al. 2003). Van Baaren et al. (2009: 2382) claim that speakers who are "more concerned with others, depend more on them, feel closer to them, or want to be liked by them, tend to take over their [conversational partners'] behaviour to greater extent". In line with this, Louwerse et al. (2012) show a social asymmetry of alignment in their data: in a map task experiment instruction followers imitated instruction givers significantly more often than the other way around.

We try to factor in a social component by looking at speaker dominance. As a proxy for dominance, we use two quantifiable factors: the number of words uttered during the conversation (*words*) and checking who is the first to label a given target object (*1_mention*). First, we counted the total number of words per interlocutor in a conversation and then calculated the relative speaker dominance[27] within that conversation. Second, for each interactional pair in our database, we annotated who introduced the topic, i.e. who was the first to label the target object talked about. Although we acknowledge that these are very coarse measures for speaker dominance, the hypothesis is that dominant speakers (i.e. the ones talking the most and the ones that are first in referring to the target object

[27] In a fictive example of one conversation where speaker1 uses 800 words and speaker2 400 words, the relative frequencies of resp. 0.67 and 0.33 would be used as values for the independent variable *words* in our database.

at hand) will align less than non-dominant speakers: the latter will be more likely to 'follow' their dominant conversational partner than the other way around.

Gaze

In the previous section (3.1) we looked at how gaze affects lexical and gestural alignment. We found an effect of SpeakerGaze (viz. the speaker fixating the face of the addressee during the prime) for lexical alignment and an effect of AddresseeGest (viz. the addressee fixating the gesture of the speaker in the prime) for gestural alignment. Because we use the exact same data set, we of course know that and how these gaze factors are relevant in explaining alignment, but we still add them to the model to measure how they relate to other potentially explanatory factors.

3.2.3 Results

BASELINE COMPARISON: INTERLOCUTORS ALIGN LEXICALLY AND GESTURALLY

Before turning to the analysis of the independent variables described above, we first want to demonstrate how we successfully tackled the content confound issue (cf. supra). In 86.8 % of the lexical interactional pairs (n=723) the interlocutors use the same word to refer to the same target object. Likewise, in 58.1 % of the gestural interactional pairs (n=536), the interlocutors use the same gestural depiction technique to refer to the same target object. In our control dataset, a set of speaker-shuffled interactions (see 3.2.2), the alignment levels are 61 % (n=1918) for lexemes and 48 % (n=1068) for gestures. The difference between the actual and the shuffled data set is significant[28] (χ^2=150.31, p<0.001 at the lexical and χ^2 =21.99, p<0.001 at the gestural level), which indicates the alignment we measure is real and not due to chance or content confound alone.

[28] In this study a baseline comparison works well because the target objects talked about are controlled for: both in the actual and the shuffled data set the interlocutors are talking about the exact same things. The only thing we manipulated in the baseline is the interactionality of the data: we omitted the temporal dependencies and turned the ordered strings of references to target objects into random strings.

DESCRIPTIVE STATISTICS: INTERACTION BETWEEN LEXICAL AND GESTURAL ALIGNMENT

Speakers who score high on lexical alignment do not necessarily score high on gestural alignment. Fig. 24a shows a scatter plot of the averaged alignment scores for lexemes and gestures per *speaker*. As is already clear from the plot, there is no correlation between the two (r=0.02, p=0.91). Likewise, when averaged across *target objects* (see Fig. 24b), there hardly is a correlation (r=-0.09, p=0.61): target objects that are often lexically aligned are not systematically gesturally aligned as well.

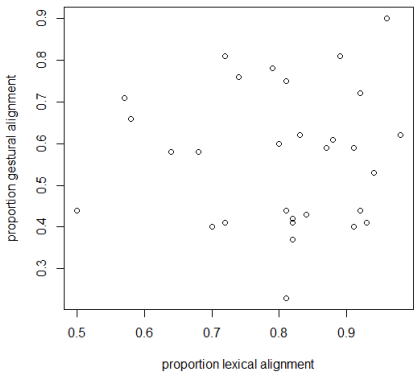

Fig. 24a: Crosstab of averaged lexical and gestural alignment per speaker

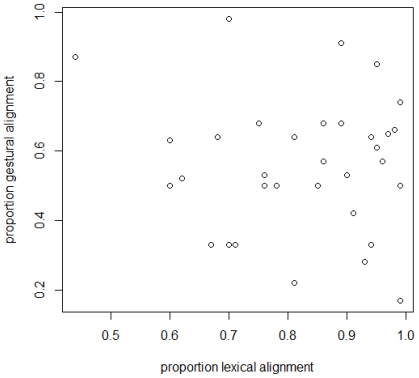

Fig.24b: Crosstab of averaged lexical and gestural alignment per object

DESCRIPTIVE STATISTICS FOR THE INDEPENDENT FACTORS

In this section we will provide a first view on the results by presenting some general trends and averages per factor. As will be clear, this descriptive statistic overview is important because it will make us rewrite some of the numerical data, for example *distance* in milliseconds, into categorical data. Before performing any statistical modelling, we want to make sure our factors are organised in the most relevant and comprehensible way. For each of the factors, we will present the results for gestural alignment (filled pattern in the graphs) and lexical alignment (checked pattern) together.

For the factor *distance*, we expected more alignment if prime and target are closer to each other. To be able to visualise alignment rates for the numerical factor *distance,* which was measured in milliseconds, we

calculated average alignment rates for 2-second intervals. Fig. 25 thus shows the average alignment rates for distances between zero and two seconds, two and four seconds, etc. For gestural alignment there is also a category *overlap*. Quite a lot (n=127) of gestural prime-target pairs occurred with a distance of 0 milliseconds, viz. prime and target gesture overlapping. For lexical prime-target pairs, there were only nine cases of overlap. Because we would have data sparseness in the overlap category at the lexical level, those nine cases were collapsed with the first 2-second interval (viz. 0-2s). Fig. 25 does not appear to hint at a linear decrease of alignment as *distance* increases, however, at the gestural level, we do see more alignment in the overlap condition (70.2%), compared to the other 2-second intervals (ranging between 44.7 and 54.3%). For gesture, *overlap* may be more crucial a factor than *distance*, and therefore we recoded the numerical factor *distance* into a categorical factor *overlap*. In the mixed effects models below, we will therefore use *distance* for lexical alignment, and *overlap* for gestural alignment.

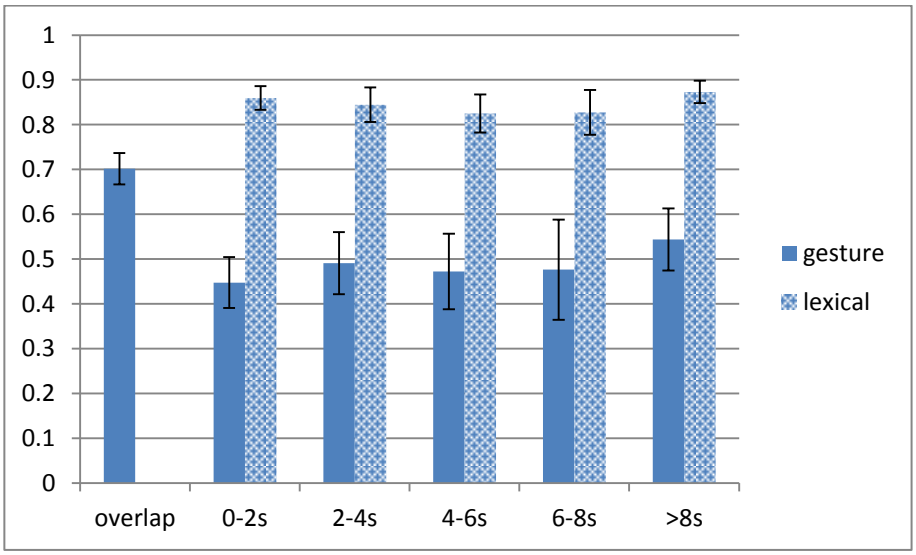

Fig. 25: Average alignment rates for 2-intervals of the factor DISTANCE *at the lexical and gestural level. Error bars indicate standard error.*

Concerning the factor *position*, we expected more alignment to occur towards the end of the conversation. It should be noted again that

conversation here stands for every discussion regarding one video animation. We calculated the relative position of each target in the prime-target pairs within each individual conversation. To graphically represent the corresponding alignment rates, we grouped the numerical factor position into four categories, resp. the first, second, third and final quarter of each conversation. Fig 26. indicates that average alignment rates for gesture are quite stable throughout the positions in the conversation. At the lexical level, however, we do see a more or less linear increase in alignment rate from the first quarter of the conversation (77.6%) to the last (91.4%). Further statistical modelling will have to provide evidence for the significance of this observation. As opposed to the factor *distance* in Fig. 26, we see no reason to rewrite the numerical factor *position* into a categorical one. For all of the statistical modelling below, we keep the relative numbers for the factor *position*.

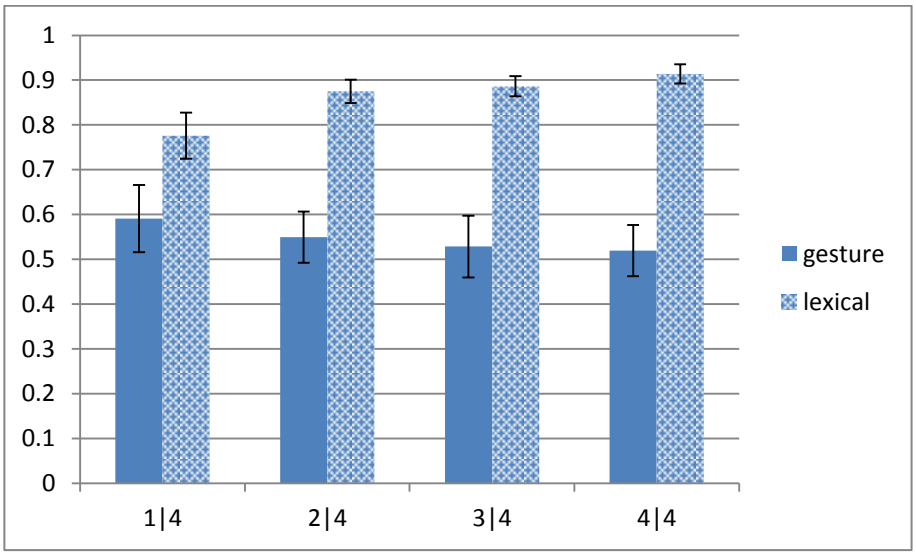

Fig. 26: Average alignment rates for the factor POSITION at the lexical and gestural level. Error bars indicate standard error.

With the factor *position* we try to check whether interlocutors align more throughout a conversation. At a higher level, we also want to check whether there is more alignment towards the end of the entire task of discussing video animations. To this end, we annotated all prime-target

pairs for the factor *block*, indicating whether the pair was performed during the first, second or third part of the task. As is clear from Fig. 27, at the gestural level we do and at the lexical we do not see an increase of alignment throughout the three blocks in the task. Further in this results section, a mixed effects model will decide on the significance of this potential effect (viz. an increase from 46.9% in the first block to 64.8% of gestural alignment in the last).

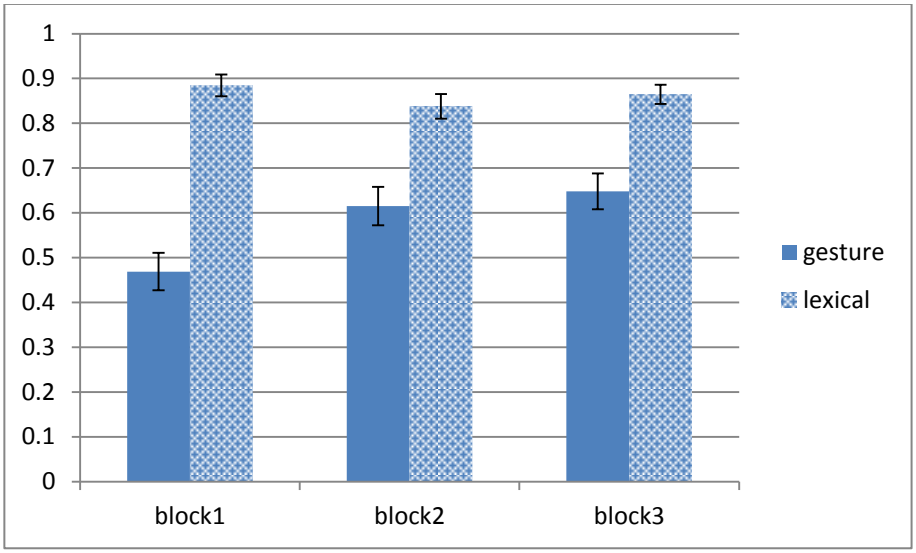

Fig. 27: Average alignment rates for the factor BLOCK at the lexical and gestural level. Error bars indicate standard error.

For the *prime* factors we wanted to check for a cumulative effect of priming, both for self-priming and for other-priming. As already indicated above, this does not mean we are looking into the difference between self-alignment and other-alignment. Our prime-target pairs all concern other-alignment because prime and target are produced by different speakers. With the factor *prime-self* we do measure whether the number of times a speaker has used a word or gesture himself, is a good predictor for aligning to another speaker. However, this crucially differs from measuring whether speakers are more likely to align to themselves, than to their conversational partner. Fig. 28 shows that both at the lexical and gestural level there is more alignment as there have been more self-primes. The effect for gesture

(ranging between 51.2% and 74.5%) appears to be more outspoken than for lexemes (84.8%-94.8%). How significant these results are, remains to be seen in the mixed effects models.

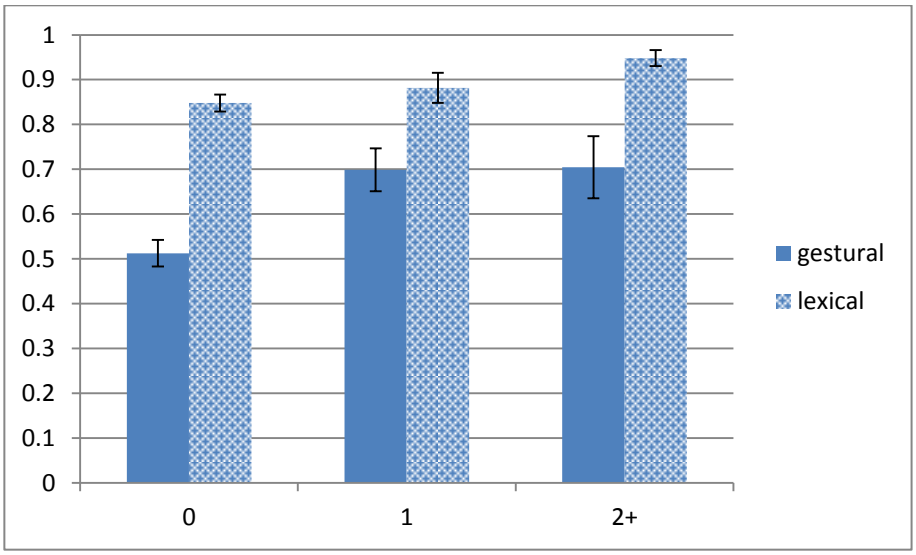

Fig. 28: Average alignment rates for the factor PRIME-SELF at the lexical and gestural level. The y-axis indicates how often, prior to the prime-target pair, the current speaker himself has already used the same word/gesture to refer to the same target object. Error bars indicate standard error.

More self-priming appears to correlate with higher alignment rates, but is there a cumulative effect in other-priming too? Fig. 29 suggests there is. Especially at the lexical level, where alignment rates hit a near-perfect 99% for cases in which, prior to the prime-target pairs, the interlocutor has already labelled the target word with the same lexeme more than two times. Note that in Fig. 29 there is a difference on the x-axis between gestural and lexical alignment. At the gestural level, the factor *prime-other* only rarely had a value of more than two (n=12). Therefore all these instances are collapsed with *prime-other* values of exactly two, creating a 'two or more' category. At the lexical level, this type of data sparseness only occurred for *prime-other* values higher than four.

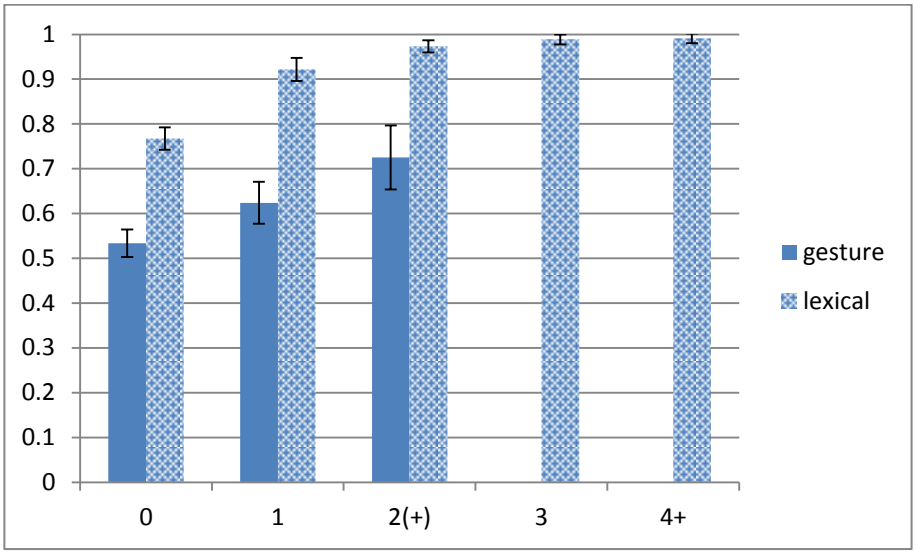

Fig. 29: Average alignment rates for the factor PRIME-OTHER at the lexical and gestural level. The y-axis indicates how often, prior to the prime-target pair, a conversational partner has already used the same word/gesture for the same target object. Error bars indicate standard error.

For the factors concerned with dominance, first, we hypothesised that speakers will align less to their partner if they themselves have introduced the target object, viz. if they were the first to lexically or gesturally label it. Fig. 30 seems to confirm this hypothesis for gestural alignment but not for lexical alignment. Second, we expected alignment rates to correlate with talkativeness. Fig. 31 shows how many words the speaker has uttered at the moment of producing the target in the prime-target pair. To visually represent this numerical factor, we used 100-word intervals. Having talked more does not seem to correlate with aligning less. This might be due to the fact that nearly all of the dyads have equally talkative speakers: only 2 dyads have number-of-words-per-conversation ratios of 0.45-0.55 or less, all the other dyads are between 0.45-0.55 ratios. This means that if one speaker in a dyad has talked a lot, the other will have too, making talkativeness not much of a distinctive factor in measuring dominance.

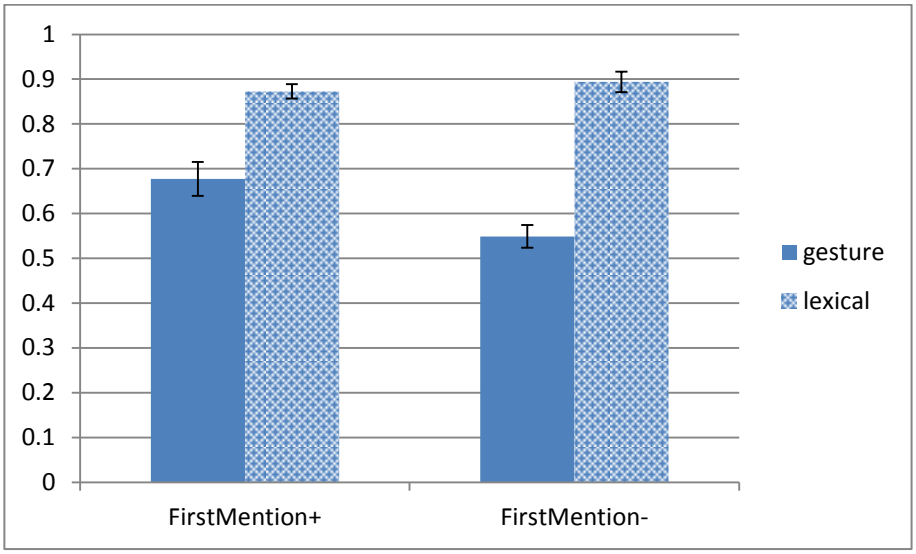

Fig. 30: Average alignment rates for the factor 1_MENTION at the lexical and gestural level. Error bars indicate standard error.

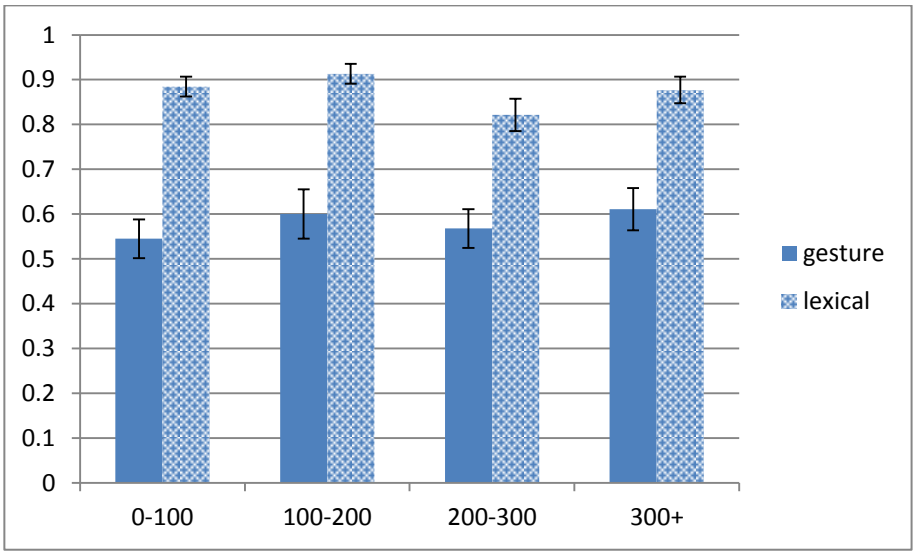

Fig. 31: Average alignment rates for the factor number of WORDS at the lexical and gestural level. Error bars indicate standard error.

What the descriptive statistics overview above shows, is that different factors seem to correlate with gestural alignment than with lexical alignment. Lexical alignment seems to correlate with position, prime-self and prime-other; gestural alignment with overlap, block, prime-self and prime-other. What has not been discussed here, but what was evident from the previous case study, is that gaze behaviour (resp. SpeakerGaze and AddresseeGest) also correlates with lexical and gestural alignment. What we want to do next is build statistical models that can uncover which of these apparently relevant factors are also significant predictors for alignment at the lexical and gestural level. To determine which fixed factors to enter in our mixed effects models, we used a forward stepwise variable selection procedure[29] and compared that output with our 'intuitive' results from the descriptive statistics section above (see Table 4). The only difference between 'intuition' and the output of the stepwise procedure concerns the factor *position* for lexical alignment (marked in grey, see also Fig. 26). In the mixed effects model we ended up using, the factor *position* was not included (cf. infra for a motivation of this choice).

Factor	**'Intuition'**		**Stepwise selection**	
	Lexic	**Gest**	**Lexic**	**Gest**
overlap	0	1	0	1
position	1	0	0	0
block	0	1	0	1
prime-self	1	1	1	1
prime-other	1	1	1	1
words	0	0	0	0
1_mention	0	1	0	1
gaze	1	1	1	1

Table 4: Overview of which fixed factors to feed into the mixed effects model

[29] We used the stepAIC function in the MASS package (Venables & Ripley, 2002). Using this function we also checked for relevant interactions between factors by calculating all possible two-way interactions, but we found none.

MIXED EFFECTS MODELS: CUMULATIVE PRIMING AS KEY FACTOR

Using descriptive statistics we have shown which factors might be relevant in explaining alignment at the lexical and gestural level. Let us now model which of these factors are significant in explaining our dependent variable *alignment*. As was done in the previous case studies (see 3.1.1 and 3.1.2), all of the mixed effects models in this results section will treat *dyad* and *object* as random factors, and gestural or lexical alignment as dependent factor. To resolve the issue of whether or not to include position as a factor in the model for lexical alignment, we first ran a linear regression analysis and then a mixed effects model with *position* included, and found it in both cases not to be a significant factor. Therefore, the reported results on lexical alignment are based on a mixed effects model without *position* as a factor in the model.

To test for collinearity issues, we calculated Pearson correlations or Cramer's V for all possible variable interactions. None of the correlation measures (for both the lexical and gestural level) were larger than 0.12, with the notable exception of *prime-self* and *prime-other.* It is not surprising that these factors are positively correlated: the longer speakers interact, the more they themselves, but also the more their partners will refer to the target objects. As the conversations unfold, and given that the two speakers in the dyad alternate in referring to the target objects, both *prime-self* and *prime-other* will increase. We will take this issue into account when discussing the results concerning these factors.

The mixed effect models for gestural and lexical alignment show that different factors predict alignment at different levels. For lexical alignment (see Table 5) we see that *prime-other* (and not *prime-self*) reached significance, as well as *gaze*. This means that the cumulative priming by the conversational partner (*prime-other*) and the gaze behaviour of that partner during the prime (*SpeakerGaze_Prime*) are good predictors for lexical alignment. To give an indication of the impact of, and the relation between the factors in the model, consider the plot in Fig. 32. This plot is based on an ANOVA analysis of the linear regression model, viz. a model without the random factors, that demonstrates how much of the explained variation can be attributed to the individual factors. It is clear from the plot that, for lexical alignment, the cumulative priming is by far the most

important factor. To evaluate the predictive power of our mixed effects model as a whole we performed two tests. First, we calculated the C-value for the model (C=0.95), and second, by comparing the fitted value for each data point to the actual value in the response variable[30] we found that the model predicts 96.5% of the data correctly. We can thus conclude that the mixed effects model for lexical alignment has predictive power, and that it clearly outperforms a *naïve* model, viz. a model that predicts every prime-target pair to be aligned (which is correct 86.8% of time because this is the average alignment rate for lexical alignment).

Fixed factor	**Estimate**	**Std.Error**	**z value**	**Pr(>|z|)**	
(Intercept)	-2.446	0.555	-4.405	1.06e-05	***
primeother	3.109	0.392	7.917	2.43e-15	***
primeself	0.052	0.178	0.291	0.77133	
SpeakerGaze	1.589	0.476	3.334	0.00085	***

Table 5: Mixed effects model for alignment at the lexical level

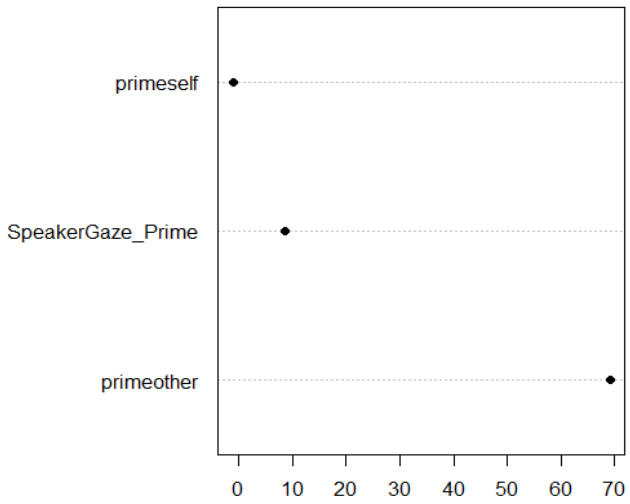

Fig. 32: Impact of the factors for lexical alignment (ANOVA)

[30] We rewrote the fitted values into a binomial dataset, with fitted values larger than 0.5 as predicting alignment (value "1"), and smaller than 0.5 predicting absence of alignment (value "0").

For gestural alignment, we see a different picture than for lexical alignment. As is clear from Table 6 all of the fixed factors in the model reached significance. However, as also demonstrated in Fig. 33, *distance* is the most important factor. This means that gestural alignment is best predicted by whether or not (and in this case whether) prime and target gesture are performed at the same time. Note that, contrary to the presentation of the data in Fig. 25, we entered *distance* as a binomial factor in the model: either prime and target gesture occurred with no distance between them (*overlap*), or there was a distance (regardless of how many milliseconds).

Fixed factor	**Estimate**	**Std.Error**	**z value**	**Pr(>|z|)**	
(Intercept)	0.073	0.348	0.211	0.83316	
primeother	0.344	0.168	2.047	0.04066	*
primeself	0.386	0.167	2.319	0.02040	*
AddrGest	0.876	0.356	2.463	0.01376	*
block	0.953	0.351	2.386	0.00678	**
distance	-1.110	0.263	-4.220	2.45e-05	***

Table 6: Mixed effects model for alignment at the gestural level

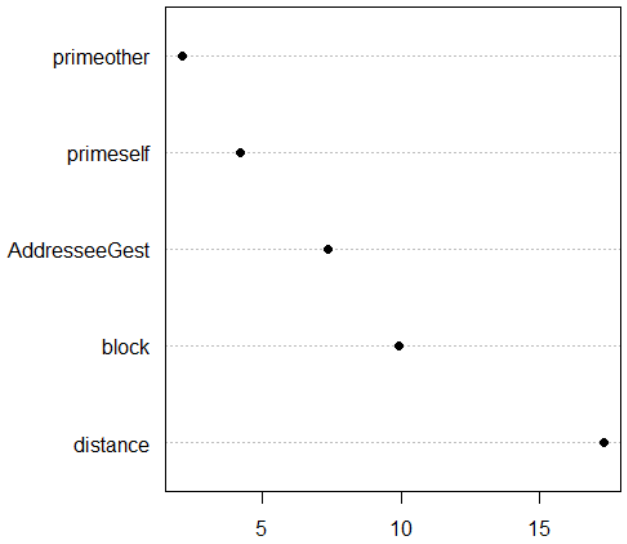

Fig. 33: Impact of the factors for gestural alignment (ANOVA)

We again assessed the overall performance of our mixed effects model for gestural alignment by calculating the C-value (0.78) and by comparing the fitted to the actual data points (67.2% of the prime-target pairs explained, compared to an overall average of 57.5% for gestural alignment). These figures are sufficient to conclude our model has explanatory and near predictive power.

3.2.4 Discussion

Existing research has shown that interlocutors match different levels of behaviour with that of their interlocutor. What separates the present study from Louwerse et al. (2012) and Bergmann & Kopp (2012) is the type of alignment under scrutiny. This study essentially deals with referential alignment (which lexical and gestural referents do interlocutors use?), as was the case in Brennan & Clark (1996) for lexical entrainment. In contrast, Louwerse et al. and Bergman & Kopp study behavioural alignment (which formal features of language use (including non-verbal) do interlocutors share?). In this study we used a uniform data design and method at the lexical and gestural level to uncover whether referential alignment occurs more frequently than chance, and to uncover by which factors it is explained.

PASSING THE BASELINE COMPARISON

Our baseline comparison test showed the alignment we measure is real and not due to chance alone. Especially at the lexical level this is an important result because the average alignment rate (0.87) is very high there. We successfully excluded that this high average occurs because speakers have only limited possibilities in lexically labelling the target objects. For example, even when talking about the abstract geometric object *circle*, interlocutors referred to it in many different terms such as “ball”, “disc”, “wheel” or “egg”. In line with Brennan & Clark (1996) we observe that lexical choice variability is high between conversations, while it is relatively low within a conversation. The different speakers in our data set use different words (and gestures) to refer to the same objects, but they tend to use the same words (and gestures) as their conversational partner. This

case study is the first to provide evidence of this type of referential alignment for gesture in face-to-face conversation.

CUMULATIVE PRIMING

If alignment (of either lexemes or gestures) were an automatic process, involving strict priming-based input-output matching, we would expect it to occur immediately from the first interactional prime-target pair, and continue ceaselessly from that point onwards. Research on social aspects of alignment (a.o. Chartrand & Bargh 1999, Hove & Risen 2009, Lakin et al. 2003, Van Baaren et al. 2009) demonstrated that a rigid automatic account of alignment cannot be maintained: alignment is clearly mediated by social factors. What this study shows, is that other (than social) contextual factors mediate alignment as well, and that priming not only constitutes an immediate but also a cumulative effect. What is more, these results were obtained in the *messiness* of spontaneous, face-to-face interaction, which underlines the methodological necessity to account for alignment in a multimodal and multifactorial way.

At the lexical level, *prime-other* was the crucial factor. At the gestural level, *prime-self* was a significant factor, albeit without accounting for a substantial part of the variation. In this vein, frequency-of-use is a stronger predictor than recency. Interlocutors (consciously or not) take into account more context than the immediately preceding utterance alone. At the lexical level, and in line with Brennan & Clark (1996), it is the accumulated behaviour of the other speaker that predicts best whether or not the current speaker will align. At the gestural level, our data support the claim made by Bergmann & Kopp (2012: 1329) that "the alignment between gestures is reliably stronger within speakers than it is across speakers", making the accumulated own behaviour the best predictor. It should be noted however, there is a crucial methodological difference (see also section 3.2.1) between this study and Bergmann & Kopp: we measure referential alignment whereas they measure formal, behavioural alignment (regardless of what the gestures refer to).

Our analysis further shows that routinisation occurs, but that it should not be read as a temporal routinisation (i.e. a process that takes some time), but rather as a referential routinisation (i.e. a process that

takes some mentions, regardless of how much time passes). The non-significant factor *position* illustrates that interlocutors do not align more as they talk longer. They only align more as they have been primed more often (by themselves or by their interlocutors). We had expected this temporal and frequency effect to coincide, but this is not the case. Apparently, the references to the target objects in our data are not evenly distributed over the conversations.

We measured no temporal effect for *position* within a conversation, but for gesture we do measure a timing effect over the three *blocks* in the task. This observation is in line with the findings of Louwerse et al. (2012: 15) who showed that "the more interlocutors interacted with each other, the more they synchronised matching behaviors with one another". Louwerse and colleagues find a temporal effect in 12 out of the 19 behaviour types under scrutiny. The behaviour types concerned with the linguistic labelling of directions, colours and digits did not show the temporal effect, whereas the gestures did. This is comparable to our finding of an effect of *block* for gestures but not for lexemes.

REFERENTIAL ALIGNMENT VS. BEHAVIOURAL ALIGNMENT

Our results show that *distance* is not a significant factor. At least, not when measured in milliseconds. For gesture, it is a significant factor when comparing the alignment of overlapping gestures with the alignment of non-overlapping gestures. This finding seems to contradict the results in Bergmann & Kopp (2012) who found a main effect of *distance* (either in terms of milliseconds or in terms of number of gestures) between prime and target gesture: the larger the distance, the less gestural alignment they measured. This difference in results can be explained by the difference in measuring the dependent variable *alignment*. In this case study we only took into account prime-target pairs that concern the same referent, i.e. the starting point was to answer the question whether and when participants use the same gesture to refer to the same object. For Bergmann & Kopp (2012) the objective was to measure formal alignment between gestures, regardless of what those gestures referred to. In their prime-target pairs, the prime might be a deictic gesture and the target a depictive gesture referring to a convertible. In this vein, if there is no

referential link between prime and target gesture, there is only a temporal link between them. It will, then, not be surprising that gestures that are further apart are less aligned. What we show is that if prime and target gesture are referentially linked, i.e. if they are expressing the same referent, the distance in milliseconds does not appear to matter in terms of gestural alignment. It is not the case that the larger the distance, the less alignment we measure. However, we do find an effect of *overlap*. This is crucially linked to the fact that we measure *referential alignment.* If participants are expressing the same content at the same time, viz. cases of gestural overlap, they are performing a gestural co-construction (cf. Kimbara 2006, 2008). Our results show significantly more alignment during these co-construction cases than during more independent gesture production.

At the lexical level our results show no significant effect for the factor *distance*, and there are not enough *overlap* cases to conduct an analysis as we did for gesture. This lack of effect for distance in lexical alignment is in line with Brennan & Clark's (1996) results, which imply "that lexical entrainment is not just a local or short-term phenomenon due to priming, but that long-term memory representations are involved." Even if prime and target are far apart, there can still be a clear alignment effect. The prominence of the factor *prime-other* further illustrates that interlocutors are not exclusively primed by very recent items: a much broader context, more specifically the effect of cumulative priming, is what appears to be governing lexical alignment the most in our data.

Cumulative priming explains a lot of the alignment measured in this study[31], but as indicated by Louwerse et al. (2012: 19) we do acknowledge that other factors that might explain (other types of) routinisation as well: "In effect, synchronisation need not be *primarily* representational: it may indicate increasingly aligned perception of the external situation". Others, such as Hove & Rise (2009) or Van Baaren et al. (2009), have demonstrated how social factors are crucial in explaining alignment. In sum, referential routinisation is driven by cumulative priming, as is shown in the present

[31] In fact, at the lexical level, cumulative priming explains more than a significant portion of the data. It explains nearly all of the data. In this vein, there is not much room left for other factors, even conversation-external factors, to account for a considerable amount of the variation in our data.

case study, but other types of routinisation might be driven by shared communicative goals, shared physical spaces, shared emotional states, etc. and not by shared mental representations alone.

GAZE

From the previous case studies (3.1.1 and 3.1.2) it was clear that gaze behaviour correlates with alignment behaviour. Addressees that are being looked at by their partner during the prime align more lexically, and addressees that have fixated the prime gesture align more gesturally. This effect of gaze appears to hold even in more complex models in which other factors are involved too. Especially at the gestural level, gaze appears to account for a fair amount of the variation (see Fig. 32).

NON-SIGNIFICANT FACTORS

The factors *words* (which speaker talks the most?) and *1_mention* (which speaker was first to introduce the target objects?) indicate that *speaker dominance* is a poor predictor for alignment. This might be due to the absence of any social hierarchy in our data. All of the participants knew each other well, they were friends and peers, and they had one common goal during the conversations, viz. to jointly try to solve the issue raised in the task ('what are the differences between the video animations for each?'). However, studies such as Louwerse et al. (2012) or Danescu-Niculescu-Mizil et al. (2012) show that if interlocutors are experimentally resp. institutionally assigned certain roles, they do show an effect of *dominance*: low power interlocutors coordinate more than high power ones. The measures for dominance in our study catch a conversation-internal type of dominance, which does not seem to generate a significant effect on language coordination as speaker *role* is shown to do. In other words, in terms of alignment frequency, it might matter more what your role is within the conversation, rather than how much you talk, or how often you introduce new topics.

3.2.5 Conclusion

There is ample evidence that interlocutors match their behaviour, both verbally and non-verbally, during interaction. In this case study we have

shown that in referring to target objects lexical and gestural alignment are predicted by different factors. Lexical alignment is predicted by the cumulative behaviour of the conversational partner, whereas for gestural alignment a range of factors are good predictors, with *overlap* being the most prominent one. Moreover, there is no correlation between gestural and lexical alignment: highly aligned speakers or target objects at the one level are not systematically highly aligned at the other.

When comparing the observations concerning referential alignment in this paper to related work on behavioural alignment in other studies, we see that *content matters*: different factors predict different types of alignment. Behavioural alignment is predicted by speaker dominance or distance and increases as the conversation unfolds, whereas for referential alignment this does not hold true. Notwithstanding the differences between the two lines of research (behaviour matching vs. conceptual pacts), the results indicate the necessity of taking into account historical facts to account for the alignment in the data. Ahistorical facts alone, or a fully mechanistic priming account alone, cannot account for the observations made in the growing body of research on multimodal alignment in conversation.

Chapter 4

Exploring the temporal dimension

In the previous chapter we tackled the phenomenon of alignment with a multimodal approach. Some of the results, however, already hinted at the importance of taking on a temporal perspective. For example, we found that overlapping gestures are significantly more often aligned than their non-overlapping counterparts, and that interlocutors gesturally aligned more over *blocks*. These results indicate that, in terms of accounting for alignment, it not only matters *what* interlocutors do, but also *when* they do it. In this chapter we will further flesh out those temporal aspects of alignment.

The shift in perspective from a multimodal to a temporal perspective entails a shift in the type of alignment under scrutiny. In the previous chapter we looked at how different factors explain alignment at different multimodal levels. The dependent variable in that chapter was a referential type of alignment. The research questions always involved which labels (either words or gestures) interlocutors used to refer to the target objects they saw in a video animation. In this chapter we include the data from the brainstorm task (see section 2.3.2), but more importantly, we no longer focus on referential alignment, but on behavioural alignment instead. This means we no longer take into account *what* people are talking or gesturing about, we only consider *at which levels* of verbal or non-verbal features of language production interlocutors formally align. In this sense, our unit of analysis shifts from prime-target pairs (of words or gestures) that refer to the same object in a video animation, to more general measures of aligned behaviour. The exact measure for alignment, and with that the measuring technique, will be dependent on the specific multimodal level and the research question at hand. For example, studying the temporal dynamics of alignment in terms of pitch will require an approach that is different from studying lexical alignment. In the latter case, we compare discrete events (i.c. words: does speaker 1 use the same word(s) as speaker 2?). In the former case, we compare between continuous data streams (i.c. values for fundamental frequency at a given sample rate).

In a first case study in this chapter we dig into the synchronisation of gaze behaviour, and answer the questions whether a participant's eye gaze is more strongly coupled to the eye gaze or to the speech of his conversational partner. In a second case study, we are not looking at

synchronisation of behaviour, but at the temporal evolution of *alignment rates* over time. For a series of gaze, speech and gesture features we answer the basic question whether participants align more (or less) towards the end of a longer conversation, and if so, whether this increase (or decrease) is gradual or not.

4.1 Case study 3: gaze synchronisation in face-to-face interaction

In this case study we want to investigate why interlocutors look at or away from each other, i.e. we want to inquire into the reason for making or breaking eye contact. People look at each other a lot during conversation, but addressees look more at their speaking partners than speakers look at their addressees (Argyle & Cook 1976, Kendon 1967). What has not yet been studied is whether interlocutors also *synchronise* their looking at each other. Do speakers look at each other at the same time? Or with a consistent time lag? And does a speaker look at a partner because that partner is looking at him? Or because that partner has started speaking? And if we do find gaze synchronisation during face-to-face interaction, will this synchronisation increase over time? Or is the synchronisation task dependent? With these questions we want to peer into both the social and the communicative functions of eye gaze.

4.1.1 Introduction & research questions

The eyes are not a mere instrument of perception, but also of production (conveying meaning about the attention, intentions, beliefs, etc. of the eyes' owner). Our eyes are well designed for vision, but also for being visible: as opposed to any other animal, the white sclera in human eyes make it possible to follow our interlocutors' gaze direction (Tomasello 2007) and read from that communicative acts such as emphasizing, disapproving or joint attention towards each other or an external object in a shared communicative setting. This dual function of perception and production makes the eyes are a unique organ. None of the other senses can pull the same trick: you cannot make yourself be heard by listening, nor does your nose smell, but you can make yourself being looked at by looking.

Although gaze has been demonstrated to serve many different functions (see introductory section on gaze alignment in 1.3.1), the exact synchronisation of eye gaze during face-to-face conversation has only received little attention. However, an understanding of the exact time alignment of gaze behaviour is critical for an understanding of the mechanisms behind it. For instance, whether an interlocutor looks at his partner even before that partner starts to speak or only when that partner tries to make eye contact, regardless of speaking events, is of importance for answering the question *why* we make eye contact. To investigate the underlying mechanisms of gaze behaviour in face-to-face conversation, in this case study we will unravel the synchronisation of a speaker's eye gaze with the gaze of the addressee and with the speech of the addressee.

Kendon (1967) and Argyle & Cook (1976) were the first to systematically look into the interaction and temporal relation between gaze and speech. They found that addressees look at their speaking partners more than the other way around; when speaker and listener switch roles (i.e. at moments of turn-taking) there nearly always is mutual gaze; speakers typically briefly look away at the beginning of their turn or during hesitations and pauses, but they do systematically look at their partner at the end of longer turns. More recent work confirmed those results (see e.g. Brône, Feyaerts & Oben 2013) and added that both verbal and non-verbal feedback markers synchronise with mutual gaze (Bavelas et al. 2002) or nuanced the results of Kendon (1967) in illustrating that turn transitions not always happen in mutual gaze, but with gaze aversion of the incoming speaker as well (Oertel et al. 2012).

More recently, not only the interaction between gaze and speech, but also the temporal relation between gaze of different conversational partners has been studied. Most studies on this topic start from a joint-attention paradigm in which participants are not looking at each other, but at a computer screen while playing a map, puzzle or matching game. Participants are reported to perform matching tasks (e.g. find a target object in a complex picture) faster if they have visual information on where their partner is looking at (Brennan et al. 2008, Frischen et al. 2007, Lachat et al. 2012, Neider et al. 2010, Richardson & Dale 2005). Also, participants

synchronise their eye movements more as they interact longer with each other (Dale et al. 2011, Hadelich & Crocker 2006).

With this case study we want to add to the literature in at least two ways. First, in the existing research on the (temporal) relation between gaze and speech, none of the studies use eye-tracking to measure the participants' gaze behaviour. Video-based estimates of eye gaze are certainly useful, but not as accurate as eye-tracking based measurements. To capture how much time interlocutors spend looking at each other, relying on video data alone could suffice. However, when studying short gaze events in their temporal relation to other gaze, gesture or speech events, eye-tracking is advisable if not necessary. Second, the studies on the (temporal) relation between gaze of different participants do use eye-tracking but do not study face-to-face interaction. The crucial difference is those studies measure gaze fixations triggered by stimuli (i.e. when are interlocutors looking at a picture on a computer screen?), whereas we will measure spontaneous, conversationally motivated fixations (i.e. when are interlocutors looking at each other?).

With the present case study then, we combine a setting of *face-to-face interactions* (with interlocutors looking at each other rather than being separated and looking at a computer screen), in which *eye-tracking* is used to measure the *synchronisation* of a speaker's gaze with the eye gaze of his conversational partner and the speech of that partner. Starting from a communicative setting that is as spontaneous as possible, we want to answer the following questions:

(i) Do interlocutors synchronise their looking at each other, regardless of their roles as speaker or addressee?
(ii) Do addressees synchronise their looking at the speaker with the speaker's speech?
(iii) Does the synchronisation measured in (i) and (ii) increase over time or differ over conversational task?
(iv) Do interlocutors adapt their gaze behaviour more to the gaze behaviour of their partner (cf. (i)) or to the speech behaviour of their partner (cf. (ii))?

4.1.2 Method & analysis

DATASET

For this case study we will use both the animation description and the brainstorm task of the Insight Interaction Corpus (see section 2.3.2 on task design in Chapter 2). By incorporating the data from the brainstorm task, we are able to check for the robustness of our findings across different interaction types. In all of the research on gaze synchronisation within the joint-attention paradigm (among others: Dale et al. 2011; Hadelich & Crocker 2006; Richardson, Dale & Kirkham 2007) interlocutors are playing a puzzle or matching game (as is the case in the animation description task). To rule out that the observed synchronisation is due to this specific type of task, we include data from another and much less restricted interaction type, viz. brainstorming. Within this animation description task we can further address research question (iv): by using the factor *block*[34] we can investigate whether there is more synchronisation the longer interlocutors are talking.

CROSS RECURRENCE QUANTIFICATION

A growing body of research (for a recent overview, see Fusaroli et al. 2014a) is using cross recurrence quantification techniques to study phenomena of behaviour matching, including the synchronisation of eye gaze (Richardson & Dale 2005, Richardson et al. 2009, Dale et al 2011). We will first demonstrate how cross recurrence analyses work and how they can be useful to our needs.

A cross recurrence analysis is a type of correlation analysis that looks for a time lag at which the overlap between two time-series is maximal. Consider the following fictive example of a 10 second interaction between two speakers (S1 and S2). The interaction between them is sampled at 1 Hz, i.e. there is one value per second. Say, in this fictive example we are looking at blinking. Red indicates a speaker has blinked during that second, green indicates there was no blink during the sample

[34] In the animation description task, participants had to discuss 15 animations in three blocks of five, because the eye-tracking system was recalibrated in between the blocks (see also case study 2 in which *block* was one of the factors in the analysis).

second. In the example below there is one moment of overlap: at second seven both speakers are blinking at the same time (indicated by the blue rectangle in Fig. 34). At this point in time they are perfectly synchronising their blinking.

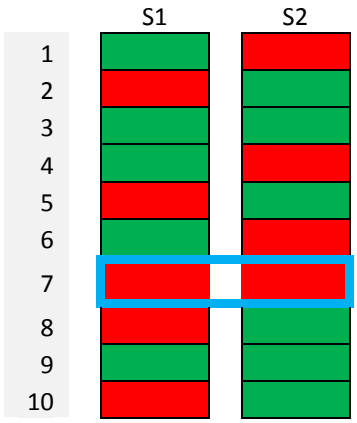

Fig. 34: Fictive example of two time-series for blinking

To study the temporal relation between the blinking of S1 and S2, it is not only interesting to measure when both speakers are blinking at the same time, it is also relevant to check whether there is a systematic time lag between the blinking of S1 and that of S2. This is exactly what a cross recurrence analysis does. When shifting the time-series of S2 one second down (see Fig. 35), we see there is a lot more overlap between the two time-series (cf. the four blue rectangles in Fig. 35). What this means is that, for this tiny example, typically S2 blinks first and S1 blinks one second later. In other words, the correlation between the two time-series is maximal at a time lag of 1 second.

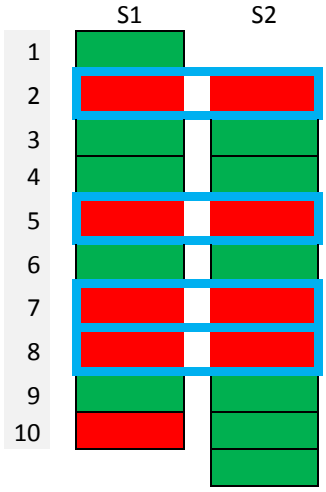

Fig. 35: Fictive example of two time-series for blinking
in which the time-series for S2 is lagged by one second to that of S1

It is of course possible to calculate the correlation between two time series for any time lag: by shifting the time-series for S2 another second down (and another and another, etc.) and by also shifting the time-series of S1 down (or that of S2 up, which is the same). In Fig. 36 we see a plot in which this calculation has been done for the fictive example above. In the middle of the plot the value on the X-axis is zero (t0). This data point indicates the amount of correlation between the actual time-series, i.e. without any lagging (cf. the visual representation in Fig. 34). For the value "1" on the X-axis, there is a clear peak: when lagging the time-series of S2 with one second, there is a lot more overlap (cf. the visual representation in Fig. 35). When, switching speaker direction, lagging the time-series of S1 with one second (this corresponds to the value "-1" on the X-axis), we do not see an increase in overlap. Basically two things can be read from this type of plots. First, the peak in the plot indicates at what time lag the correlation or overlap between two time-series is maximal. Second, the position of the peak relative to t0 indicates who follows who: to the right of t0 we measure the correlation or overlap for S1 following S2 (as in Fig. 35); to the left we see S2 following S1. For the fictive example here we can read from the plot in Fig. 36 that S1 typically follows S2 (and not the other way around) at a

time lag of one second. In other words, typically one second after a blink by S2, also S1 blinks.

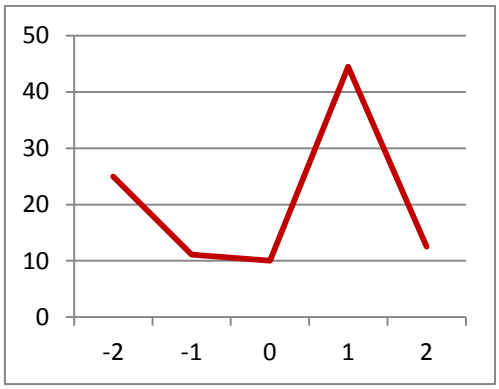

Fig. 36: Simplified cross recurrence plot for the fictive example in Fig. 34 and 35

DATA PREPARATION

As already mentioned, for this case study we will use both the data from the animation description and the brainstorm task in the Insight Interaction Corpus. All of the transcriptions were done in ELAN (Lausberg & Sloetjes 2009). Because ELAN works with a waveform viewer that visualises the amplitude of the audio signal, it was possible to very precisely annotate the on- and offset of speech for each of the participants. To allow for a cross recurrence analysis, the existing transcriptions were sampled (10 Hz) into categorical time-series: every 100 ms we polled the transcription tiers and scored, per participant, whether there was speech ("1") or not ("0"). Completely analogous to our speech coding, also for gaze we created categorical time-series from the existing gaze annotations with a sample rate of 10 Hz. For gaze we coded "1" if there was gaze towards the face of the interlocutor and "0" if there was not.

It is important to stress that for this case study we only take into account whether or not participants are speaking, and whether or not they are looking at each other's face. The result of sampling both the speech and gaze annotations can be schematised as follows:

Gaze (S1)	1	1	1	1	1	1	0	…
Gaze (S2)	1	1	0	0	0	0	0	…
Speech (S1)	0	0	0	0	0	0	0	…
Speech (S2)	1	1	1	1	1	1	1	…
	100	200	300	400	500	600	700	TIME (MS) …

SOFTWARE & EXAMPLE ANALYSIS

To measure how eye gaze is synchronised in our data, we used the R-package developed by Coco & Dale (2014) to perform cross recurrence quantification analysis (CRQA) on our data. Fig. 37 shows an example of such a CRQA of gaze for one dyad in the animation description task. The Y-axis indicates the recurrence rate; the X-axis represents a time scale in seconds with t0 in the middle. The values on the Y-axis are difficult to interpret as such. Because they are dependent on the frequency of the phenomenon (i.c. looking at the face of the conversational partner) and on the sample rate, they do no not make intuitive sense. What matters in interpreting the plot, are the peaks and their position relative to t0.

In the example in Fig. 37 we see a double bell curve, indicating there is synchronisation of looking at each other's face. Moreover, this synchronisation is asymmetrical: speaker 1 follows speaker 2 more often (higher peak in the plot) and faster (peak is closer to t0 on the horizontal axis) than the other way around. The synchronisation we observe for this dyad (red line) is higher than the synchronisation we would expect to occur by chance (blue line, cf. infra for the calculation of a baseline).

In the example in Fig. 37 we looked at synchronisation of gaze by speaker 1 and gaze by speaker 2. As was outlined in the research questions, we will also be looking at the synchronisation of gaze and speech. Because both gaze and speech are sampled at the same rate and in the same categorical way, a CRQA between them is methodologically identical to the CRQA-plot in Fig. 37 (i.e. a gaze-gaze synchronisation). The type of plot in Fig. 37 was performed for each of the dyads, in both the animation description and brainstorm task, for the levels of comparison sketched in the research questions (synchronisation of face fixations with face fixations and synchronisation of face fixations by the addressee with speaking of the

speaker). In the results section we will present averaged plots, based on the individual CRQA plots such as the one in Fig. 37.

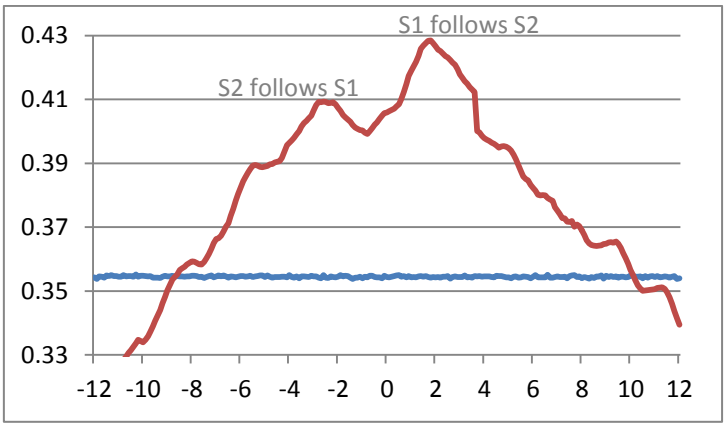

Fig. 37: CRQA-plot of two speakers in the animation description task for "looking at the face of the partner" (x-axis: time lag in seconds; y-axis: recurrence rate)

BASELINE COMPARISONS

In the previous chapter on referential alignment, we checked whether speakers use words/gestures because their conversational partners just did, or because of chance (e.g. because there simply are no alternative for those words/gestures). In this case study we want to do the same thing and check whether the synchronisation we measure in the CRQA plots occurs because people are actually adapting their behaviour to one another. To ensure that our results were not obtained by chance, we computed a baseline, following among others Richardson & Dale (2005) and Louwerse et al. (2012). The time-series of the gaze and speech data (see section on data preparation) were disarranged, creating a random distribution of behaviour across time rather than a conversationally motivated one. Using this procedure we created 1.000 pairs of temporally randomised speech and gaze data. On each of those pairs we then performed a CRQA. The average of those cross recurrence analyses should be read as the chance level of synchronisation: only if the CRQA plot of the actual data is above the averaged baseline plot, the synchronisation is real and not due to chance alone. An example of this was already shown in Fig. 37. The blue line in that plot represents the baseline obtained by this randomisation procedure. The

red line is the CRQA plot of the real data. Whenever the red line rises above the baseline, we can say that actual synchronisation occurs.

Apart from the time-randomisation routine, we also tested a speaker-randomisation. For this analysis we coupled the time series of every speaker to every other speaker in the corpus (i.e. creating fake dialogues between interlocutors that never actually interacted with each other), and performed CRQA to those pairs of time series. The baselines resulting from a speaker-randomisation were all lower than those obtained by time-randomisation. To lower the risk of observing synchrony where there is none, we only use the baseline data obtained by the time-randomisation procedure: because our statistical tests are based on the difference between real data and baseline data, we only report and use the most conservative, i.e. the highest, baseline.

4.1.3 Results

SYNCHRONISATION WITH EYE GAZE

The interlocutors in our corpus synchronise their eye gaze during face-to-face conversation, or to be more correct, they synchronise their looking at each other's face. We already saw an example of this type of synchronisation for one dyad in Fig. 37. When averaged across dyads and interaction *type* (i.e. animation description and brainstorm task), we find that gaze in face-to-face conversation is indeed strongly synchronised (see fig. 38 for the averaged CRQA plot).

To test for the significance of the difference between the gaze synchronisation in the real data (fig. 38, red line) and the baseline data (fig. 38, blue line), we calculated a mixed effects model. As fixed effect we entered the variable *real-vs-base* (binomially indicating whether the data come from the shuffled baseline or the real interactions). As random effect we added *dyad* (categorically indicating a code for each dyad) to the model. The recurrence rates (i.e. the values on the vertical axes in the individual cross recurrence plots, such as the one in Fig. 37) were the dependent variable. The mixed effects model confirmed what is in fact already very clear from the plot in Fig. 38: the real CRQA plot in red differs significantly from the baseline plot in blue ($t=57{,}81$, $p<0.001$), reliably indicating that interlocutors synchronise their looking at each other.

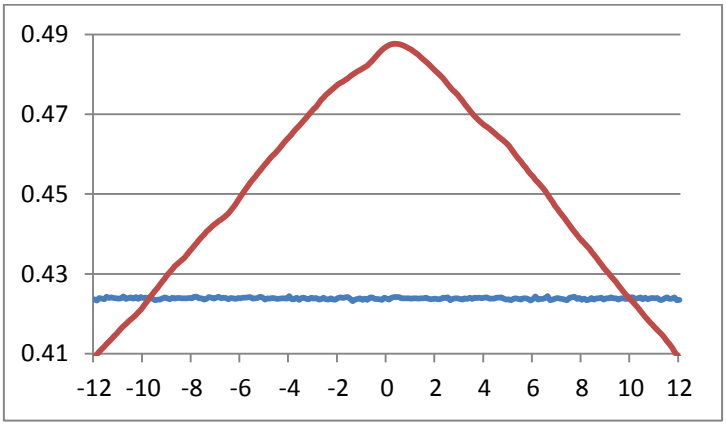

Fig. 38: Averaged CRQA plot of gaze at face by both interlocutors (x-axis: time lag in seconds; y-axis: recurrence rate)

Before we continue with the rest of the results, we first want to clarify one important methodological issue. When performing CRQA, it is necessary to set a lag window. In principle it is possible to calculate recurrence rates for time lags up to tens of seconds or even minutes. However, stretching the time lags too far increases the risk of observing synchronisation that is not due to the interaction. For eye gaze, for example, it is not reasonable to claim that two gaze shifts that are three minutes apart are in any way linked to each other. On the other hand, reducing the time lags to, say, only one second, increases the risk of missing synchronisation that is actually there: two gaze shifts that are two seconds apart may very well be instances of genuine synchronisation.

Therefore, for all the statistical analyses we report in this case study, we use a dynamic window of analysis. This window is bound by the crossing of the real recurrence rates with the baseline recurrence rates. For the results in Fig. 38 this means, in our mixed effects model, we only compared the real data to the baseline data for time lags ranging from -9,7 to +9.9 seconds. Again, this window of analysis is defined by the crossing of the recurrence plot for the real data and the baseline data. Time lags outside of this window are not considered to be synchronisation, and therefore left out of the analyses. This window of analysis will be different for each statistical analysis we make, and for the rest of the results section

we assume that the data that go into the mixed effects models pertain to the dynamic window as described above.

SYNCHRONISATION WITH SPEECH

Speakers not only synchronise their eye gaze behaviour with their interlocutor's eye gaze, but also with their interlocutor's speech. Fig. 39 shows the averaged CRQA plot for fixations at the face of the interlocutor and speech by the interlocutor. A mixed effects model confirms this synchronisation differs significantly from the baseline synchronisation ($t=52.56$, $p<0.001$). As can be observed from Fig. 39, and opposed to the synchronisation of gaze with gaze of the interlocutor in Fig. 38, the peak of the recurrence plot does not coincide with t0. The peak here is at 0.3 seconds. This means that the gaze signal needs 0.3 seconds to be maximally aligned with the speech signal. In other words, typically 0.3 seconds after a speaker's speech, an addressee is fixating his speaking partner.

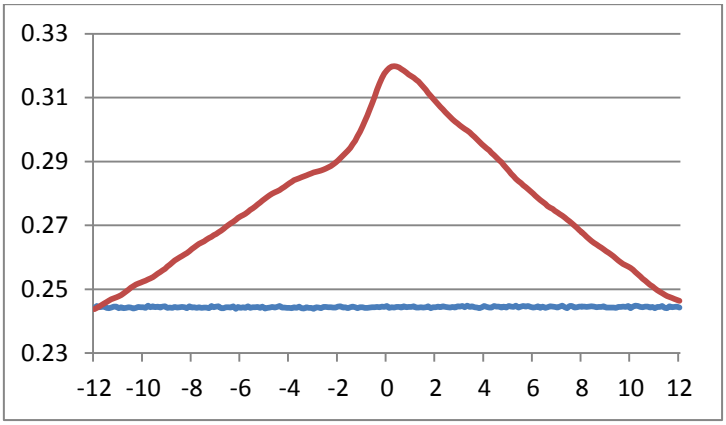

Fig. 39: Averaged cross recurrence of gaze at face by one interlocutor and speaking by the other interlocutor (x-axis: time lag in seconds; y-axis: recurrence rate)

SYNCHRONISATION INCREASES OVER *BLOCK*

So far, we provided evidence that interlocutors in face-to-face interaction synchronise their gaze behaviour with the gaze behaviour of their partner and with the speech behaviour of their partner (as compared to a time-randomised baseline). In line with what Dale et al. (2011) found for synchronisation of eye gaze and Louwerse et al. (2012) for synchronisation

of a range of other phenomena, we found further proof that gaze synchronisation is affected by conversation time. The longer interlocutors engage in communication, the more synchronisation we observed. Fig. 40 illustrates there is more synchronisation in block 2 and 3 than in the first block. As was explained in the section on how CRQA works, the recurrence rates on the y-axis are dependent on the frequency of the phenomenon. Therefore, it would not be fair to directly compare the recurrence values of the three blocks. The fact that the peaks in Fig. 40 for block 2 and 3 are higher than that of block 1, might be due to the fact that there simply are more or longer fixations on the face of the partner in those blocks, regardless of their synchronisation. To exclude that we are measuring frequency of the phenomenon instead of synchronisation, we normalised the data by calculating the difference between the real recurrence value and the baseline recurrence value (i.e. we created a baseline per block) for each data point. A mixed effects model, with the normalised recurrence rates as dependent, *block* as fixed effect and *dyad* as random effect variable shows that gaze synchronisation indeed significantly increases as *block* ($t=24.51$, $p<0.001$) increases.

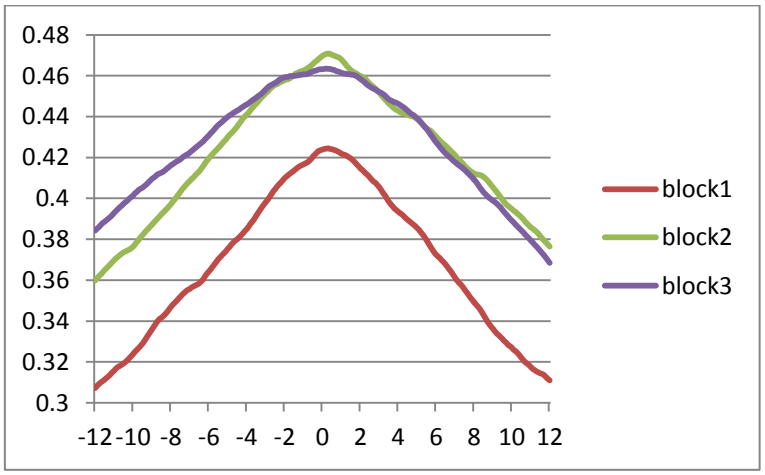

Fig. 40: Averaged cross recurrence per block of gaze at face by both interlocutors in the animation description task (x-axis: time lag in seconds; y-axis: recurrence rate)

Not only does the synchronisation of gaze with the gaze of the interlocutor increase over time, also the synchronisation of gaze with speech does. Fig.

41 demonstrates how the cross recurrence rates of fixations on the face of the partner and speech of the partner unfold over *blocks*. A mixed effects model (again with normalised recurrence rate as dependent, *block* as fixed and *dyad* as random factor) shows that this synchronisation clearly increases over the three *blocks* (t=31.74 p<0.001).

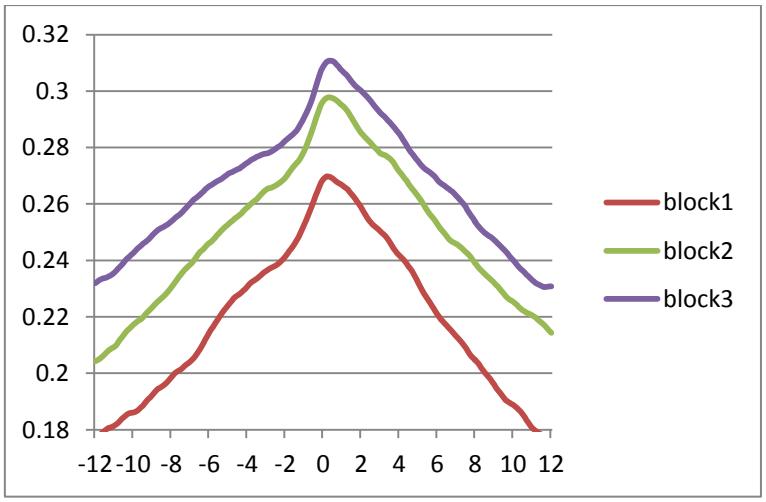

Fig. 41: Averaged cross recurrence per block of gaze at face by one interlocutor and speaking by the other interlocutor, in the animation description task (x-axis: time lag in seconds; y-axis: recurrence rate)

SYNCHRONISATION IS TASK-DEPENDENT

To check whether the conversational task affects gaze synchronisation, we compared the synchronisation in the animation description task to that in the brainstorm task. Mixed effects models with normalised recurrence rates (cf. infra) as dependent variable, *dyad* as random factor and *task* as fixed factor reveal that both for synchronisation of gaze with gaze (t=12.47, p<0.001) and for the synchronisation of gaze with speech (t=21.71, p<0.001) there is a significant difference between the animation description task and the brainstorm task. For both types (gaze-gaze and gaze-speech) there is significantly more synchronisation in the former than in the latter task, proving the task-dependency of the synchronisation under scrutiny.

SYNCHRONISATION WITH GAZE IS STRONGER THAN WITH SPEECH

If we do not compare conversational *tasks* but synchronisation *types*, we see that eye gaze is more closely coupled to eye gaze than to speech. More specifically, we calculated a mixed effects model with the normalised recurrence rates (cf. infra) as dependent variable, synchronisation *type* (either gaze-gaze synchronisation or gaze-speech synchronisation) as fixed factor and *dyad* as random factor. Eye gaze appears to be significantly more strongly coupled with the eye gaze of the other participant than with the speech of the other participant ($t=14.09$, $p<0.001$).

4.1.4 Discussion

The findings in this study can be summarised as follows:

(i) Interlocutors synchronise their eye gaze with that of their partner. More specifically, if a speaker is looking at his partner, that partner will look back (and vice-versa).

(ii) Interlocutors time-align their eye gaze with their partner's speech. More specifically, addressees synchronise their fixating the speaker with the speaker's speech.

(iii) The synchronisation in (i) and (ii) is affected by *conversation time* (more synchronisation towards the end of the animation description task) and by the type of *task* (more synchronisation in animation description than brainstorm).

(iv) Interlocutors synchronise their eye gaze more with their partner's gaze than with their partner's speech.

With result (ii) we have replicated the observations of Kendon (1967), Argyle & Cook (1976) and more recently Bavelas et al. (2002) and Cummins (2011): addressees typically look at their partner while listening. However, we were also able to refine those observations and dig into the exact temporal coupling of gaze and speech behaviour: addressees look at their partners typically 0.3 seconds after that partner is speaking. This type of synchronisation could be read as an action-reaction or stimulus-response type of behaviour. The speaker starts talking and the addressee pays attention to what the speaker is doing by looking at him. In this sense, the time lag of 0.3 seconds might be a residue of the processing on the part of

the addressee that a speaker has started talking. However, because we measure the gaze-speech synchronisation at the most general level (e.g. making abstraction of the contents, duration or embedding in turn management of the speech), many different phenomena might underlie the time lag we observe[35]. Further research is needed to confirm such a time lag and to address the issue of *why* it occurs.

For the synchronisation of gaze with the gaze of the interlocutor, we found a recurrence rate maximum at t0. This perfect synchrony indicates that interlocutors 'know' their partner is going to look at them, before it actually happens. If they did not model and therefore project/predict their conversational partner's next eye gaze move, they would not be able to respond with a -literally- instantaneous gaze event themselves. In this view, it is a result of mutually taking into account each other's gaze behaviour (cf. infra). Taking the results in (i) and (ii) together, this either implies that interlocutors are better at predicting when they will be looked at, than when their partner will start talking, or that it is more relevant to answer a face fixation with a face fixation than it is to answer speech with a face fixation.

In this case study the observed gaze synchronisation occurs in a face-to-face setting. This is crucially different from studies like Richardson & Dale (2005, 2009), Richardson et al. (2007) or Dale et al. (2011). What we measured is not a synchronisation of gaze towards a (shared) external object, but a synchronisation of gaze towards each other. In this respect, the coupled eye movements are a borderline case of *synchronisation* and could be regarded as *coordination*: speakers are not looking at the same thing at the same time (i.e. there is no imitation, no focus on the same item) but they are doing comparable things (i.e. looking at each other, each

[35] For example, from conversation analysis there is evidence that speakers often look at their conversational partner *before* that partners starts talking (Mondada 2007). This phenomenon in which a speaker's next turn is projected by the gaze behaviour of his conversational partner, cannot be confirmed (nor excluded) by our results. We do find a time lag in the opposite direction (i.e. addressees fixate the speaker's face typically 0.3 seconds after he is talking), but the presence of speech (our unit of analysis) and the presence of turns (Mondada's unit of analysis) are too disparate to be directly compared.

focussing on a different item). Because gaze has both a perception and production function, gaze synchronisation in a face-to-face setting starts a perception-production loop: speaker 1 knows that speaker 2 knows that speaker 1 knows that [ad infinitum] they are looking at each other. This is not the case in studies that start from a non-face-to-face puzzle solving task, where speaker 1 does not know where speaker 2 is looking at[36]. The fact that in the latter set of studies, participants' eye gaze behaviour does synchronise, is indicative of the communicative function of synchronisation. The coupled looking at the same item makes referential sense in the whole of the communicative act of puzzle solving. In our study the gaze synchronisation does not make referential sense and hints at the result of a basic grounding process, a low-level social function. It is certainly not (only) a residual of the cognitive efforts to solve a task, but a core social catalyst that underpins the interaction itself. In this sense, it reminds us of Clark's most basic action ladder (1996: 147) of getting an "addressee to attend to the message" as a prerequisite for successful communication (cf. also the notion of *mental connection* in Wheatley 2012: 593). Regardless of any propositional content, a first and crucial step is to signal your partner your intentions of setting up a joint project and have that partner consider those intentions. Synchronising eye gaze might be a means of climbing this first rung of the action ladder. The observation that gaze is significantly more synchronised with gaze than with speech and that interlocutors appear to be able to predict when they will be looked at (because they typically look at their partner at exactly the same time), underlines the importance of this social function of gaze synchronisation in face-to-face interaction.

In line with Dale et al. (2011), Hadelich & Crocker (2006) and Louwerse et al. (2012) we found that both gaze synchronisation and gaze-speech coordination are dependent on the time spent interacting: interlocutors synchronise more as they interact longer. This has been linked to the social function of eye gaze described above: social affiliation and

[36] There are some studies (Frischen et al. 2007, Brennan et al. 2008, Neider et al. 2010, Lachat et al. 2012) that demonstrate that knowing where your conversational partner is looking, results in faster or better task completion. Again, this alludes to the referential, communicative function of gaze, as opposed to the grounding, social function we observe in our data.

synchronisation mutually feed into each other (Chartrand & Bargh 1999, Van Baaren et al. 2009, Hove & Risen 2009), making it likely that more synchronisation occurs as interaction time unfolds. This interpretation, however, does not apply all that much to our specific data set because the interlocutors all know each other really well. Their 'liking each other' is long established, and not likely to relate to what happens during our experiment. A different interpretation is linked to the type of interaction (i.c. a collaborative task) in which *conversation time* was a significant factor. Maybe interlocutors synchronise their gaze behaviour more because they develop a routine in solving the task. In this sense the underlying mechanism is not a social but a communicative one. Because the task is repeated (in this case study 15 times), it might be unfair to equal the factor *block* to unfolding of time throughout discourse. This unfolding is different from conversations in which there is no repeated task, so the temporal effect we measure might not be a clean case of time elapsing, but rather of repeatedly performing a comparable task. Linked to this interpretation is the fact that we found significantly more synchronisation in the animation description task than in the brainstorm task. Together, the effect of the task type and the possible effect of task repetition, raise a methodological issue: are the results better explained by task-related factors than by the social, cognitive or interactional factors we actually envisage? This issue also applies beyond this study. Nearly all of the psycholinguistic research on the topic of gaze synchronisation involves eye gaze during very comparable tasks (maze games, map task, etc. in which subtasks are repeated). This urges the question how far results from a task-based interaction can be generalised to any type of interaction.

In the introduction we stated that gaze has a multitude of functions. Although gaze behaviour can be considered as being constrained (Richardson et al. 2007, Richardson & Dale 2009, Fusaroli et al. 2014b) by a lot of factors such as common ground between the speakers, visually salient items in a conversational setting, speech behaviour of the conversational partner, etc., gaze synchronisation is a prominent feature of face-to-face interaction. Our measuring a lot of gaze synchronisation, which runs through any of the possibly very local, ad hoc or idiosyncratic functions gaze might have, is remarkable. Regardless of what interlocutors

say, even regardless of what they see, they time align their looking at each other and their listening with looking at their partner. This complete abstraction of the contents of the conversation makes the results remarkable, but it also raises questions for future research. How independent is the synchronisation we measure? To what extent is the synchronisation dependent on what people say? Or how they say it? Or who says it? Maybe we synchronise our eye gaze in face-to-face interaction more to people we like, or to messages we deem relevant?

What this study has demonstrated, and in this respect it links up with the results in the first case study, is that eye gaze serves social and communicative functions at the same time. We came to this conclusion by zooming in on the temporal relations of gaze behaviour and speech behaviour. In a next case study we will further show the relevance of exploring the temporal dimension of coordinative behaviour.

4.2 Case study 4: a temporal account of speech and gestural alignment

In the previous case study we looked into one type of temporal relation between behavioural events, viz. synchronisation. In the present case study we address a different kind of temporal relation: the amount of alignment as a function of time. This temporal issue was already briefly touched upon in case study 2. There we demonstrated that interaction time affects alignment rates at some multimodal levels. More specifically, we showed that the factor *block* significantly correlates with gestural alignment: there was more gestural alignment towards the end of the experiment. This temporal effect related to referential alignment. In the present case study we will not only study the alignment of referring expressions in speech and gesture, but also of other formal properties of the speech and gesture signal. Furthermore, we will look into the temporal aspect in a more fine-grained manner than only factoring in *block*, which roughly split up the data in three large chunks. In sum, we will study the temporal dynamics of different formal aspects of speech and gestural alignment in detail. Do interlocutors, at different multimodal levels, align more towards the end of a conversation? And does this possible increase arise gradually? Or does it occur in local peaks and spurts?

As already explained at the beginning of this chapter, with this case study we are not interested in the temporal relation between prime and target in an interactional pair. For gaze (Dale et al. 2011, Hadelich & Crocker 2006, case study 3 in this dissertation), and for a range of other multimodal levels (Louwerse et al. 2012) it has already been demonstrated that *synchronisation* increases over time. Crucial in these analyses was how much time there is between behaviour of S1 and of S2, and whether there was a consistent time lag between that behaviour. In this case study, however, we are not concerned with how much time there is between prime and target, but in how much alignment (i.e. how many aligned prime-target pairs) we measure throughout a (longer) conversation.

In the literature on alignment so far, hardly any attention has been paid to these temporal dynamics of alignment rates. One exception is a small body of research in the domain of prosodic alignment (De Looze et al. 2014, Kousidis et al. 2009, Vaughan 2011). To be able to compare our

method and our results to those studies, we include some 'low hanging fruit' features of intonation, viz. fundamental frequency, loudness and speech rate in our analysis. Apart from comparing our results with existing studies, we also want to dig into the temporal dynamics of alignment at other multimodal levels, viz. the lexical, syntactic and gestural level. Because each of those specific levels of analysis requires its own method, we will report on them separately, rather than first present all the different measuring techniques in one section and then describe all of the results into another section.

All of the analyses in this chapter are based on the data from the brainstorm task. This was done for two reasons. First, and linked to the discussion section in the previous case study, routinisation through repetition of the task is not an issue in these data. Any temporal effect we might find, will not be due to the top-down imposed structure of the task, but to the bottom-up dynamics of the interaction. Second, the interactions in the brainstorm task are significantly longer (between 4 and 12 minutes) than those in the animation description task (hardly ever longer than 2 minutes). Since our method and research questions are aimed at longer stretches of discourse, also in this sense the choice for the brainstorm task is obvious.

4.2.1 The temporal dynamics of gestural alignment

The temporal aspect of gestural alignment we already touched upon in case study 3 concerned the time difference between prime and target in a prime-target pair: overlapping gestures (i.e. without any time difference between prime and target) are significantly more aligned than non-overlapping ones. For the present study we want to know whether, regardless of this time difference between prime and target, interlocutors align their gestures more as they interact longer. As was done in case study 3, we will focus on a specific type of gesture, viz. representational gestures.

When studying gestural alignment, an obvious but difficult issue is to set formal criteria as to what counts as gestural alignment. If speaker 1 points by using one extended finger, and speaker 2 points by using three extended fingers, are those two gestures then aligned? If speaker 1 shows his fist to indicate he is very angry, and speaker 2 briefly clenches his fist

during a subtle beat gesture, again, are those two gestures aligned? As a gateway into this difficult issue, parallel to what we did in case study 3, we will focus on the representation technique (following the typology of Streeck 2008: 292-295) of the depictive gestures in the brainstorm task (n=378). The representation technique is reliably measurable and suited to answer the holistic question 'are these two gestures alike?'[37]. A crucial difference with case study 3 is that we not only consider referential alignment in the present study, i.e. we disregard whether prime and target are expressing the same referent. For example, if speaker 1 is performing a drawing gesture to refer to a cell phone, and speaker 2 uses a drawing gesture to refer to a hand bag, we will consider those two gestures to be aligned even though they are not referring to the same thing[38].

METHOD & ANALYSIS

Maybe needless to repeat, but it is crucial to stress that, as opposed to the previous case study on the temporal dimension of gaze (see 4.1), in the present study we are interested in *alignment* and not in *synchronisation*. In other words, it matters whether prime and target gesture are produced with the same representation technique, but it does not matter how much time there (typically) is between prime and target. Therefore, cross recurrence techniques, such as the one we used in section 4.1, do not apply to this study. Inspired by methods for measuring the temporal dynamics of alignment of intonation, we used a technique of time-aligned moving averages (TAMA, see De Looze et al. 2014, Kousidis et al. 2009) to map the amount of gestural alignment throughout time.

[37] Annotating gesture always involves some kind of reduction to the original gesture. We do acknowledge that it would be relevant to incorporate gesture features such as hand shape, trajectory, place in the gesture space, duration, etc. but to fit in the scope of this dissertation we decided to restrict ourselves to the happy medium of measuring alignment at the level of representation technique.

[38] In the animation description task, we designed an experiment to elicit repeated references to the same object. This was not possible in a more free task such as brainstorming. Only comparing gestures with identical referents would not yield sufficient cases for a sensible analysis.

Before explaining how we used TAMA to analyse the temporal dynamics of gestural alignment[39], let us first zoom in on how we defined gestural alignment in this study. Consider the following example that illustrates our definition of gestural alignment.

begintime	**endtime**	**gest-S1**	**gest-S2**	**alignment**
54813	55807	point		
65170	70980	handle		
70980	73457	point		
75480	76954		shape	
83057	83963		point	aligned
84543	85336		model	
85336	86251		bound	
88373	90552	handle		
92185	93330	handle		
93390	94196		point	
95782	96588	handle		
96588	98200	model		aligned
97868	98584		handle	aligned
98270	100646		model	aligned
103037	111180		model	
111200	118577	handle		aligned
118735	121457		handle	aligned
121894	123802		point	
125396	129643	handle		aligned
135797	137642	shape		
137642	141920	handle		

Table 7: Example excerpt of gesture annotation in the brainstorm task

The example above (Table 7) is an excerpt of the gesture annotation for one of the dyads in the brainstorm task. For each gesture we see the begin and end time (in milliseconds), the representation technique and whether or not the gesture is aligned according to our procedure. This procedure

[39] All of the calculations of alignment cases and all of the TAMA analyses were done by writing scripts in Perl.

involves three parameters. We consider a prime-target pair to be aligned if prime and target:

(i) have the same representation technique;
(ii) fall within a 40 second range;
(iii) occur in an interactional pair.

Parameter (i) is straightforward, but by means of illustration, consider the three still images below (Fig 42a-c). They correspond to the first three gestures that involve *modelling* as representation technique (see Table 7). In these three cases the interlocutors are modelling a cell phone they are talking about by representing the phone with their flat (left) hand.

Fig. 42a

Fig. 42b

Fig. 42c

For parameter (ii) we use the same window of analysis as we will do throughout this case study, viz. a window of 40 seconds. We do acknowledge this window is rather arbitrary, but we chose it for matters of consistency throughout the temporal analyses at different levels (cf. infra), and because we do not want to overstretch the window of analysis[40]. This means we do not want to consider, for example, any two pointing gestures

[40] On the one hand, De Looze et al. (2014) and Vaughan (2011) in their studies on prosodic alignment use a window size of 100 and 200 seconds, respectively. Large windows like that would produce overly smooth plots, and not allow to peer into alignment in a sufficiently fine-grained manner. On the other hand, Kousidis et al. (2009), also for prosodic alignment, use a 20 second window, which would underestimate the alignment ratio in our data.

that are three minutes apart as still being aligned. In defining a fixed window of analysis we risk missing out on instances of alignment that are actually there (e.g. two identical gestures that happen to be 50 seconds apart) rather than, the other way around, we risk counting a lot of noise in our data. However, because we always compute a baseline[41] to compare our real results to, we keep ourselves from drawing undesirable conclusions: if using a 40 second window of analysis under- or overgenerates alignment cases, we will also under- or overgenerate the alignment in the baseline.

The third parameter (iii) in our procedure to measure gestural alignment is related to our measuring technique in case study 2 (see Fig. 17, p. 92): we only consider interactional prime-target pairs. Consider the first three pointing gestures in Table 7: first speaker 1 uses two pointing gestures, then speaker 2 uses a pointing gesture. We do not count this as two instances of gestural alignment, but only consider the interactional pair of the second and third pointing gesture. This means we only take two gestures to be aligned if they are adjacent, i.e. if no other gesture of the same representation technique occurs in between them.

Now that we have defined how we measure gestural alignment, we can explain how we proceeded with the TAMA-technique (De Looze et al. 2014, Kousidis 2008) to map the temporal dynamics. Crucial to the TAMA-method is to slide a *window of analysis* over the time-series data. As is shown in Fig. 43 (in between green brackets) we first create a window of analysis of 40 seconds. This windows covers the first 40 seconds of the conversation. Within this window we calculate the amount of gestural alignment. This amount is the ratio of aligned gestures (cf. supra, see also Table 7) over the maximum amount of possibly aligned gestures within the given window. We then shift the window of analysis further across the time axis, with a *step* of five seconds (yellow brackets in Fig. 43, a window from second 5 to second 45), calculate the same type of alignment ratio for the gestures in that window, and so on. This way we have a rate for gestural alignment every five seconds throughout the entire conversation. As a

[41] For all of the analyses in this case study, we again created a set of 1000 fake dialogues by time-randomizing the actual annotations.

result, the plot in Fig. 39 can be read as the unfolding of gestural alignment over time.

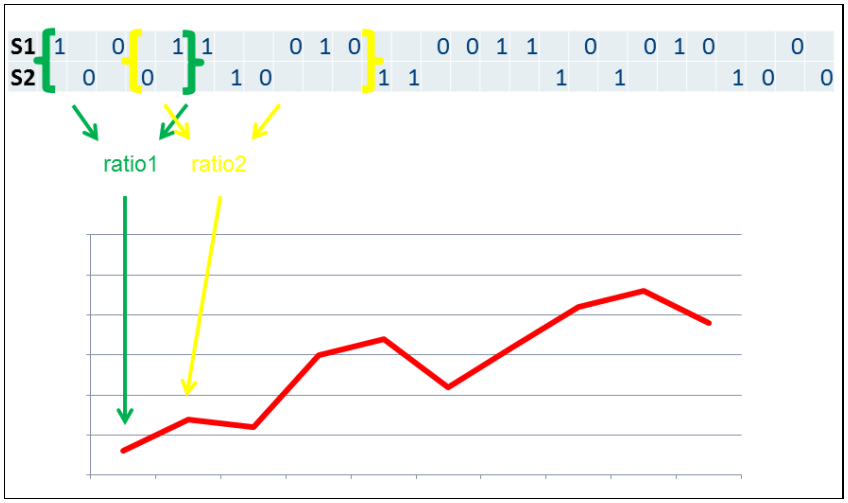

Fig. 43: Fictive example as illustration of TAMA method

Because we are working with averages within rather small windows of analysis, we risk that the data points in our TAMA plots are based on very few instances. If there are only two gestures within a window, and they are aligned, it would be unfair to claim there is 'more' alignment than in a window in which 18 out of the 20 gestures are aligned. To avoid basing the alignment ratios on too little instances, we only computed a ratio for windows in which there are at least five possible aligned pairs of gestures. Fig. 44 provides an example of such a TAMA plot for one of the dyads in the brainstorm task.

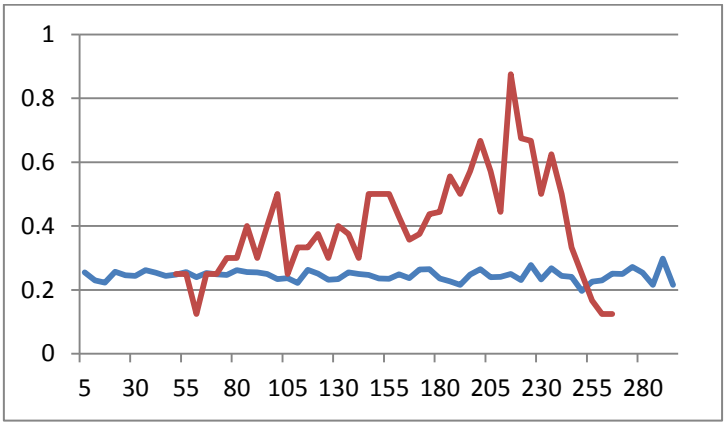

Fig. 44: TAMA plot for gestural alignment for a dyad in the brainstorm task
(x-axis: conversation time in seconds; y-axis: alignment ratio)

For the dyad in Fig. 44 we see there is an insufficient amount of gestures in the first and the last part of the interaction (no data points for the red line). In between we see a slight increase in gestural alignment as the conversation unfolds. When we apply this TAMA method to all of the data in the brainstorm task, we observe quite some missing data points due to an insufficient number of gestures. In total, 32 per cent of the data points in the TAMA plots are blank because of this self-inflicted issue of data sparseness. In fact, the plot in Fig. 44 is one of the few uninterrupted TAMA plots in the corpus. Analysing individual TAMA-plots is therefore difficult, however, averaging across dyads can still be indicative of the overall temporal dynamics of gestural alignment.

RESULTS

Although we had some issues concerning data sparseness, each data point in the TAMA plot in Fig. 45 (which is averaged across dyads) is based on data from at least eight different dyads. This enables a fair interpretation of what is clear from Fig. 45: interlocutors systematically align their representational gestures during the brainstorm task, and they align more as they interact longer with each other. A mixed effects model with the variable *real-vs-base* as fixed effect, *dyad* as random effect and the gestural alignment ratios as dependent variable, confirmed the difference between real and baseline data is significant ($t=9.53$, $p<0.001$). A comparable mixed

effects model with interaction time (in seconds) as fixed effect, indicates the increase in alignment we see in the plot is also statistically significant (t=8.48, p<0.001).

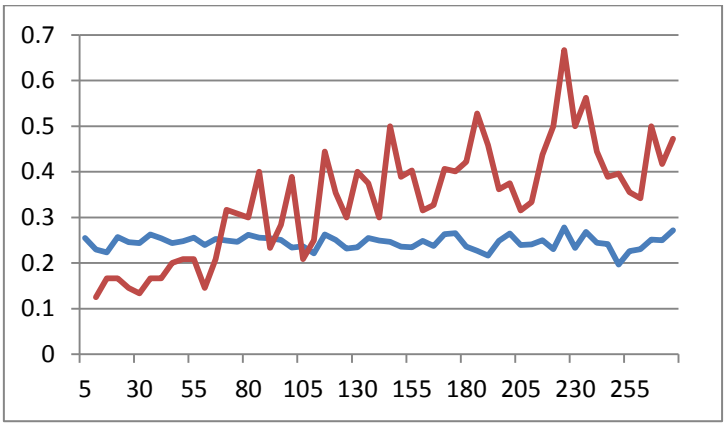

Fig. 45: Averaged TAMA plot for gestural alignment in the brainstorm task (x-axis: conversation time in seconds; y-axis: alignment ratio)

DISCUSSION

Research on temporal aspects of gestural alignment, so far, only focussed on the time difference between prime and target gesture (as in e.g. Bergmann & Kopp 2012). To the best of our knowledge, no research has studied the amount of gestural alignment across conversation time[42]. Both in the animation description task (see Fig. 27 in the previous chapter) and in the brainstorm task (see Fig. 45 above), we see an increase of gestural alignment over interaction time. What the present study indicates, is that this increase of alignment is not only due to an increase of task difficulty or the building of task-solving routines (an issue we discussed in the previous study on gaze synchronisation). Also regardless of difficulty and routine building, gestural alignment appears to increase purely as a function of time. Because of the relative data sparseness, it is not possible to look at

[42] Louwerse et al. (2012) do study the factor *time* in relation to gesture but (i) they study gesture *synchronisation*, which is a different phenomenon than gestural *alignment* and (ii) they factor in time by studying synchronisation *over* repeated tasks, whereas we study the impact of the factor time *within* longer conversations.

the individual TAMA plots per dyad, and make any claims as to how gradual or suddenly this increase comes about.

The participants in the brainstorm task all spontaneously shaped the features and functionalities of the cell phone in gesture. The increasing recurrence of depiction type might hint at an increasingly joint construal of the brainstormed ideas. Maybe interlocutors' gestures become more aligned because their representations of what the cell phone, specifically branded for women, should look like. However, future research and an in-depth analysis of the conversation (as in Kimbara 2006) should point out whether this interpretation is plausible.

4.2.2 The temporal dynamics of speech alignment

The TAMA method explained and deployed in the previous study on gesture has been used for studying prosodic alignment (De Looze et al. 2014, Kousidis et al. 2009, Vaughan 2011). These authors found that alignment of different prosodic features such as loudness, pitch and speech rate "dynamically evolves over phases of conversations rather than increases/decreases continuously over the course of a conversation" (De Looze et al. 2014: 20). Using the TAMA method, these studies indeed demonstrate that prosodic alignment changes dynamically over the course of individual conversations, rather than increasing or decreasing gradually. However, what these studies do not show, is whether there is a trend of increase or decrease of alignment of intonation, when averaged across conversations. Furthermore, no systematic inquiry has been done into the temporal dynamics of alignment at the speech level, outside the domain of prosody. Therefore, in the present study, apart from pitch, loudness and speech rate, we will also focus on how lexical and syntactic alignment (dynamically) varies as a function of interaction time. Because each of the levels of the speech signal under scrutiny (i.e. prosodic, lexical, syntactic) in this section 4.2.2 requires a different methodological approach, we will report on them in separate subsections.

THE TEMPORAL DYNAMICS OF PROSODIC ALIGNMENT

When people interact, they copy each other's prosodic features. This phenomenon has caught a lot of attention over the past decades (among

many others: Collins 1998, Giles et al. 1991, Gregory & Hoyt 1982, Stanford & Webster 1992, Webb 1972). Some researchers found that prosodic alignment increases over time. This was demonstrated both within very strict experimental settings of shadowing tasks[43] (e.g. Fowler et al. 2003), as in more natural settings of targeted collaborative tasks (e.g. Pardo 2006). Recently, researchers also studied prosodic alignment in spontaneous conversations, however, they failed to find evidence of an increase in alignment rates. Instead, they found prosodic alignment to be a highly dynamic, rather than gradual phenomenon (De Looze et al. 2014, Kousidis et al. 2009, Vaughan 2011). With the present study we want to add to the existing literature in two ways. First, nearly all of the research on the dynamics of prosodic alignment (both in experimental and corpus based research) is based on conversations in which participants are unable to see each other. We want to test whether the same results hold true for face-to-face conversations. Second, apart from looking into dynamic temporal patterns for individual conversations, we want to check whether, when averaged across conversations, we can distinguish a more global pattern of prosodic alignment. In what follows we first discuss the method and results for alignment of speech rate, followed by that of pitch and loudness.

Alignment of speech rate

For all of the transcriptions in the Insight Interaction Corpus, the exact on- and offsets of speech were very precisely anchored to the time axis (see section on data preparation in 4.1.2). This allows us to accurately measure how fast participants speak. As already explained for gestural alignment (cf. supra) for all of the analyses on the temporal dynamics of alignment, we will use a *window* of analysis of 40 seconds and a *step* of 5 seconds. This means, for every speaker separately, we first calculate the average speech rate for the first 40 seconds of the interaction, then for second 5 to 45, then for second 10 to 50, and so on. Fig. 46 is an example of a TAMA plot for speech rate for one of the dyads.

[43] In shadowing tasks participants typically get to hear a syllable, word or word group, and are then asked to repeat as fast as possible. Researchers then study the phonetic similarity between prime and response.

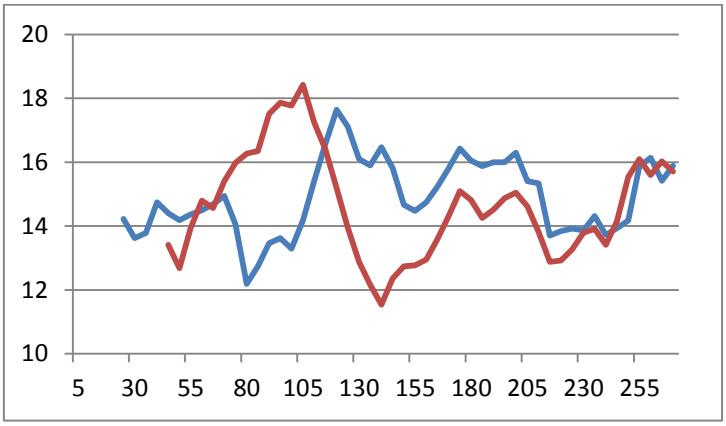

Fig. 46: TAMA plot for speech rate of speaker 1 and speaker 2 in one conversation of the brainstorm task (x-axis: conversation time in seconds, y-axis: speech rate)

In the TAMA plot above (Fig. 46), the y-axis represents the speech rates of both participants in terms of number of transcription characters per (spoken) second[44]. In the first place, the plot demonstrates that speakers vary their speech rate across time. Second, roughly from second 155 onwards, the speakers converge in terms of speech rate. From a methodological point of view, note that not for every *step* we have a speech rate value: at the beginning of the conversation, there are blanks in the data. Just as for gestural alignment (cf. supra) this is linked to the issue of data sparseness. In calculating the TAMA plots, we incorporated a rule stating that for every 40 second window of analysis, there has to be at least 5 seconds of speech before we calculate a speech rate value. Because we are working with ratios, we deemed it unfair to base a value for speech rate on very short segments. If a speaker in a 40 second window only says "uhu that's right", we want to be careful and not plainly extrapolate the speech

[44] We chose this measure of speech rate to avoid time-consuming, extra annotations (e.g. in terms of syllables). We acknowledge it is not a common measure of speech rate (as number of syllables per minute is), but to us the speech rate measure itself is not important. Only the comparison between speakers, and the comparison with a time-randomised baseline is. Moreover, for one dyad we calculated TAMA plots for both measures of speech rate (i.e. syllables per second and transcription characters per second) and found they hardly differed (Pearson's r of 0.96 for speaker 1; 0.98 for speaker2).

rate of this very short segment to the entire 40 second window. In contrast to the analyses at the gestural level, we only rarely encountered this data sparseness issue for speech rate (viz. in 3 % of the windows).

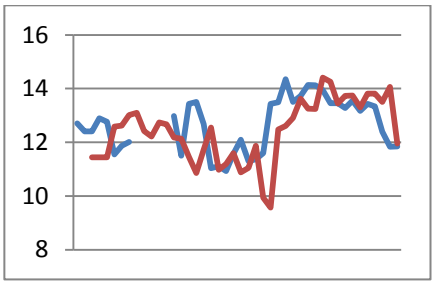

Fig. 47a: Parallel TAMA plots

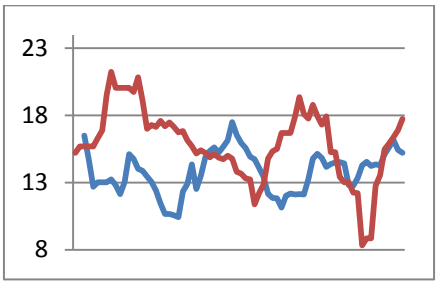

Fig. 47b: Unrelated TAMA plots

When applying the TAMA method to the entire data set, we see that some speakers' plots are almost parallel (Fig. 47a), some are converging (Fig. 46) and some seem to be totally unrelated (Fig. 47b). However, for this study we are not as much interested in individual conversations. We want to uncover which systematic patterns appear when comparing across dyads. More specifically, we want to check whether alignment of speech rate increases over time. Such an increase over time could be manifested in two ways:

(i) speakers converge in absolute terms, i.e. at the end of the conversation they are talking at roughly the same speech rate (whereas at the beginning they were not).

(ii) speakers converge in relative terms, i.e. at the end of the conversation their individual TAMA plots are more strongly correlated than at the beginning. This does not imply they are talking at the same speech rate (as in (i)), only that they adapt their speech rate to that of their partner in relative terms: if speaker 1 speeds up, also speaker 2 does and if speaker 1 slows down, also speaker 2 does.

To check for an increase of speech rate alignment over time as in (i) we calculated (the absolute value of) the difference between the speech rate of speaker 1 and that of speaker 2 for every data point in the TAMA plot. The resulting plot reflects the difference in speech rate in absolute terms

across time. If we then average across conversations (see Fig. 48), we see this difference fluctuates but not systematically decreases or increases over time. This means that speakers do no converge (or diverge) in terms of absolute speech rate. In other words, speakers do not come to talk at the same speech rate as the interaction unfolds.

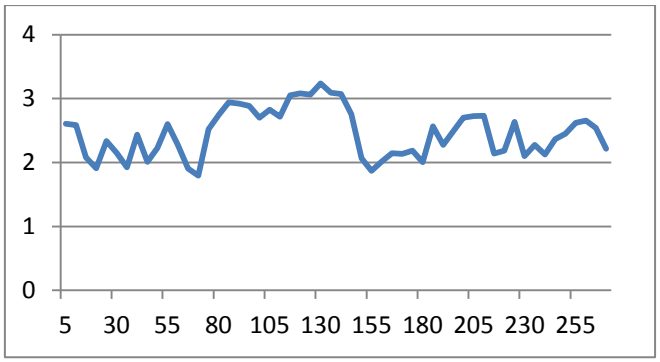

Fig. 48: Difference in speech rates between speaker 1 and speaker 2 over conversation time, averaged across conversations

To look into the question raised in (ii), i.e. to check how speakers align their speech rates in relative terms, we used the following procedure. First, we normalised the individual TAMA plots (as the ones in Fig. 46 and 47) by computing z-scores. This allows us to better compare across conversations, i.e. to better compare across different speech styles. Some speakers vary a lot in terms of speech rate; others more or less maintain a constant speech rate. Some participants speak very fast, others very slow. Because we are interested in the question whether speakers adapt their behaviour to that of their partner, regardless of whether they end up talking at the same (absolute) speech rate, normalising the data provides more reliable results. Second, after having computed z-scores, we again calculated TAMA plots for all speakers. Third, to measure how interlocutors align in terms of speech rate, we calculated the correlation between the two (z-score based) TAMA plots for every dyad. Because we are interested in alignment of speech rate over time, we did not just calculate a correlation coefficient for the entire conversation, but we again applied the TAMA principle and calculated correlation coefficients throughout time. This means, again with

a window of analysis of 40 seconds and a step of 5 seconds, we first calculated a correlation coefficient for the data in the first 40 seconds, then for second 5 to 45, then from second 10 to 50, and so on. The resulting plot indicates the amount of correlation (i.e. alignment) between the two TAMA plots for speech rate throughout time. Fig. 49 is an example of such a correlation plot, in which we see a correlation analysis based on the TAMA plots for speech rate in Fig. 46. As was already clear from Fig. 46, the interlocutors start aligning their speech rate systematically, somewhat halfway into the conversation. This is also reflected in the correlation plot below in Fig. 49.

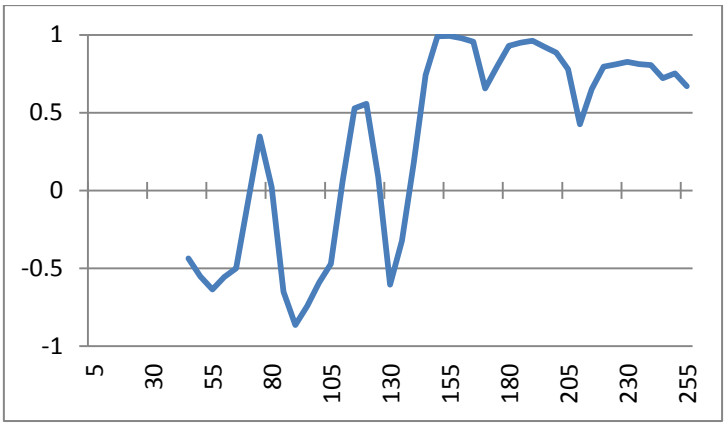

Fig. 49: Correlation of speech rate across time for one dyad in the brainstorm task (x-axis: interaction time in seconds, y-axis: correlation values)

When we average the correlation plots (like the one above in Fig. 49) across all interactions, we get the overview in Fig. 50. In this plot we see the averaged correlation values across time (in red) and a baseline[45] (in blue). Fig. 50 provides evidence that speakers do not align their speech rates more as the interaction unfolds. What is more, a mixed effects model (with *dyads* as random effect and *real-vs-base* as fixed effect) reveals that, on

[45] As was done throughout this chapter, the baseline was obtained by time-randomizing the (z-score based) TAMA plots for speech rate a thousand fold, applying the correlation analysis to all those shuffled data sets, and averaging across those correlation analyses. Not surprisingly, the baseline is more or less equal to a correlation with value "0", i.e. the baseline represents random noise.

average, interlocutors do not adapt their speech rate to that of their partner (t=3.97, p=0.07).

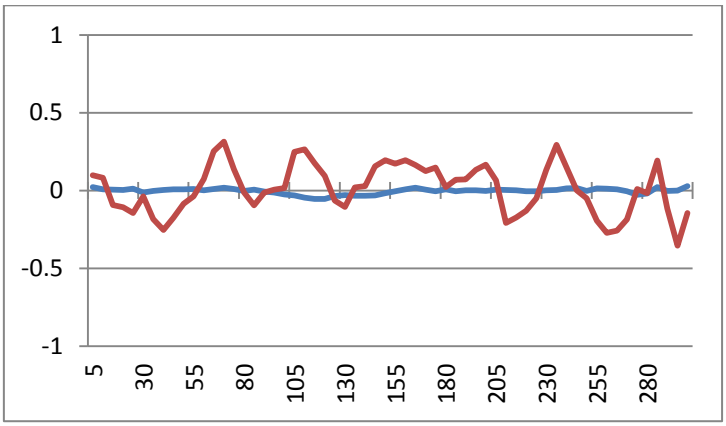

Fig. 50: Correlation of speech rate across time averaged across dyads (x-axis: interaction time in seconds, y-axis: correlation values)

Taking the results in this section together, we see that speakers locally align in terms of speech rate, but not globally. This ties in with what De Looze et al. (2014) and Bonin et al. (2013) found for conversations via telephone. The results suggest that speech rate is a rather stable property of speech production, a strongly speaker-tied phenomenon that is not easily subject to change during face-to-face interaction. As suggested by Bonin et al. (2013) changes in speech rate might be more difficult to perceive than changes in e.g. pitch or loudness. Because interlocutors are less sensitive to speech rate changes they might align less at this level. Moreover, as demonstrated in the overview article in Juslin & Sherer (2005), participants are better at assessing emotions drawing on pitch and loudness, than drawing on speech rate. If speech rate is much less linked to (recognising) emotional states, this might also lead to less alignment at this level, compared to pitch and loudness (cf. infra).

Although we did not find a temporal pattern for speech rate alignment, and we did not even measure significantly more overall alignment than in a shuffled baseline, we still take home one relevant methodological achievement. Studying the temporal dynamics of speech

rate alignment requires specific techniques. Without our TAMA approach we would not have been able to even observe the local peaks of speech rate alignment that are definitely there (as for example the clear case of convergence in Fig. 46/49). In itself these local patterns are difficult to interpret, but when comparing the patterns at one level (speech rate) to their counterparts at another level (lexical alignment, for example) the TAMA method will demonstrate its usefulness (cf. Chapter 5).

Alignment of pitch

Interlocutors appear to locally, but not globally, align their speech rates. In this section we want to find out whether the same observation applies to another prosodic cue, viz. pitch. For this study we define pitch in terms of averaged fundamental frequency (f0). This means, we are not looking into pitch dynamics, i.e. alignment of specific pitch movements. We want to answer the more general question of whether interlocutors lower or rise their pitch when their conversational partner does. In the previous section on alignment of speech rate, we explained our method of measuring the temporal dynamics of the phenomenon. Because that method is directly transferrable to this section on pitch alignment (and also to the next sections on loudness, lexical and syntactic alignment), we will be able to move to the results more quickly.

To check for local and global patterns of pitch alignment we first used Praat (Boersma & Weenink 2009) to calculate the f0 values for all of the speech in the brainstorm task. This was done by writing a script that, at a sample rate of 10 Hz, returned an f0 value for all of the voiced speech segments. To avoid taking very low intensity sounds (such as background noise or the humming of the video camera or laptops) for human voices, we set an intensity threshold: sounds that had an intensity value of less than five per cent of the loudest sound were not considered for f0 analysis. Depending on the gender of the speaker, we also set a pitch threshold to avoid unlikely outliers. For men, we only considered f0 values in a range between 50 and 400 Hz. For women, this was between 100 and 500 Hz. After automatically calculating the f0 values in Praat, we manually checked the data for outliers, and we omitted all of the values in which the two participants were speaking at the same time. Even though speakers were

being recorded with separate microphones, there was too much interference to reliably measure pitch for the speakers separately during such moments of overlap.

After manually checking the data, we calculated TAMA plots: for each speaker we first averaged the f0 values for the first 40 seconds of the interaction, then for second 5 to second 45, then from second 10 to 50, and so on. Parallel to what we did for speech rate, we only computed an average f0 value if there was at least 5 seconds of speech during the 40 second window of analysis. This was an issue for only 5% of the data.

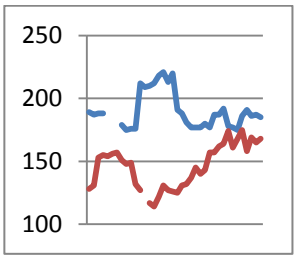

Fig. 51a:Converging plots

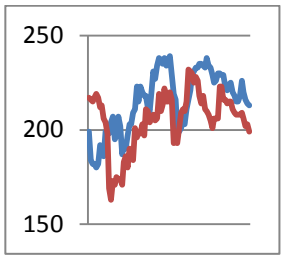

Fig. 51b:Parallel plots

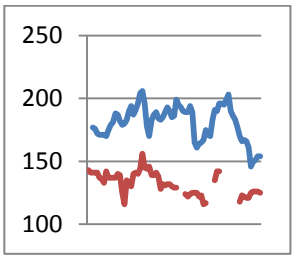

Fig. 51c: Unrelated plots

As is clear from Fig. 51 some dyads converge in terms of average pitch (as in Fig. 51a), others run in neat parallel (as in Fig. 51b) and some appear to be totally unrelated (as in Fig. 51c). By calculating the absolute value of the difference between the pitch value of speaker 1 and that of speaker 2, and by averaging that result across dyads (cf. supra), we found that speakers do not converge in terms of absolute f0 values. In other words, participants do not come to speak at an average frequency closer to that of their partner as the interaction unfolds.

To measure the temporal dynamics of pitch alignment, we first normalised the individual TAMA plots by calculating z-scores. On the basis of those data we then computed correlation plots, again with a *window* of analysis of 40 seconds and a *step* of 5 seconds, and averaged those correlation plots across dyads (cf. supra). Fig. 52 shows the results of that analysis, including a baseline based on a thousand time-randomised, fictive TAMA plots.

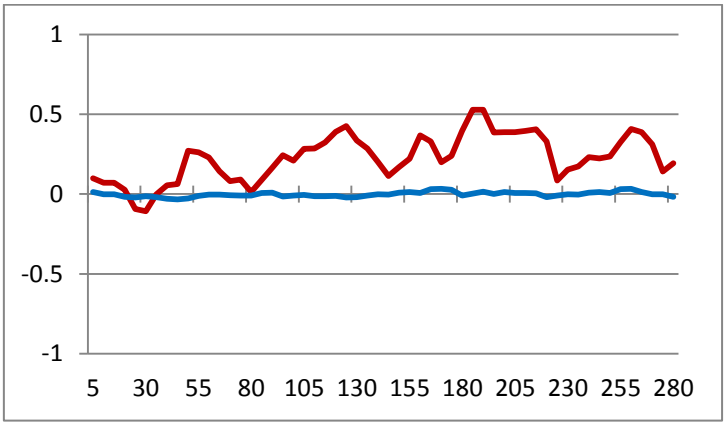

Fig. 52: Correlation of pitch (f0) across time averaged across dyads (x-axis: interaction time in seconds, y-axis: correlation values)

What Fig. 52 demonstrates is that not only do speakers globally align in terms of average pitch, there also seems to be a trend of increasing alignment over interaction time. A first mixed effects model, with *dyad* as random effect and *real-vs-base* as fixed effect, confirms that the real data differ significantly from the baseline data ($t=10.82$, $p<0.001$). A second model, with *conversation time* as fixed effect, further proves that the pitch alignment increases as conversation time increases ($t=3.01$, $p=0.02$). This effect, however, is rather small and only on the verge of being significant.

What our analyses demonstrate is that pitch alignment not only varies dynamically (i.e. in local peaks), but that it also varies systematically (i.e. it increases over time). Related studies, such as Bonin et al. 2013, De Looze et al. 2014 and Kousidis et al. 2009, used a comparable method and found neither a global nor an increasing effect. They only report local peaks of pitch synchronisation. Because the related studies all involve participants talking over the phone, what the differing results might imply, is that face-to-face contact facilitates pitch synchronisation. Given there is a link between pitch and emotion (recognition), as described by Juslin & Sherer (2005), we can hypothesise that face-to-face conversations allow for the full multimodal repertoire of expressing emotion. This could lead to more interpersonal affinity and affect, compared to conversations over the telephone, which in turn might lead to less pitch alignment in phone calls than in face-to-face speech.

Note that what separates Bonin and De Looze's studies from the present study is the interaction *setting* and not the interaction *type*. With this we want to make clear that any possible difference in results between Bonin and De Looze's studies and our study is not due to the fact that participants in our study are performing a task rather than having an entirely free conversation. Bonin et al. (2013) also use a task (i.c. the "Winter Survival Task", see also Kousidis et al. 2009), whereas De Looze et al. (2014) study spontaneous conversations in which participants receive no task at all. Bonin and De Looze's interaction *type* differs (targeted collaborative task vs. spontaneous speech) but their interaction *setting* is identical (separated speakers talking over the phone). Hence, possible differences in results with the present study can be linked to differences in interaction *setting* (i.c. face-to-face vs. physically separated), not interaction *type* (task based vs. free conversation).

Alignment of loudness

Over the course of the brainstorm task we saw that speakers increasingly align their average pitch. Now, do we see a comparable increase for loudness as well? To answer that question, we followed the exact same procedure as we did for pitch alignment. This means we are -again- not looking for specific patterns of emphasis or rhythm. We are only interested in average loudness in terms of decibels (dB) because we basically want to answer the very general question of whether interlocutors speak up or quiet down when their conversational partners do. To reliably measure loudness we had Praat (Boersma & Weenink 2009) poll the audio files at a 10 Hz sample rate. As we did for pitch, we made sure we were only measuring the loudness of actual speech (and not of background noise) by setting an intensity threshold (cf. supra). We also manually checked for outliers and omitted cases of overlapping speech. The corrected output from Praat was then subjected to the TAMA method: for each speaker we first averaged the dB values for the first 40 seconds of the interaction, then for second 5 to 45, then from second 10 to 50, and so on. Again, parallel to the previous studies on speech rate and pitch, we only computed an average loudness value if there was at least 5 seconds of speech during the 40 second window of analysis.

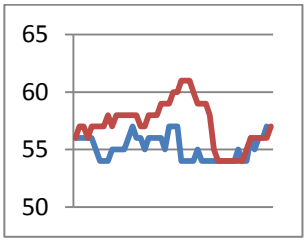

Fig. 53a: Converging plots

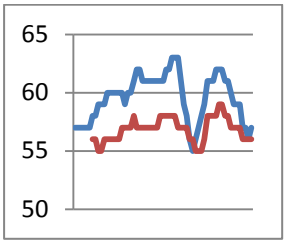

Fig. 53b: Parallel plots

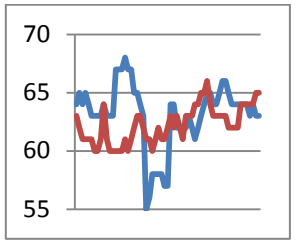

Fig. 53c: Unrelated plots

The individual TAMA plots are analogous to those of speech rate and pitch alignment: some dyads converge at the end of the conversation (see Fig. 53a), other plots are nearly parallel (as in Fig. 53b) and still other plots are nearly unrelated (as in Fig. 53c). Measuring whether speakers converge in terms of absolute loudness, i.e. whether they come to produce speech at an equal decibel-level, was not possible for this study. Because during the recordings we did not precisely control for the distance between mouth and microphone or the angle of the microphone towards the speakers' mouths, it would be unfair to compare the absolute values in decibel we obtained by our script in Praat.

To measure the temporal dynamics of alignment of loudness, we first calculated z-scores to normalise the original data, and then calculated correlation plots with a *window* of analysis of 40 seconds and a *step* of 5 seconds (cf. supra). What we see in Fig. 54 is an average (across dyads) of those correlation plots (red line) and the baseline comparison (blue line).

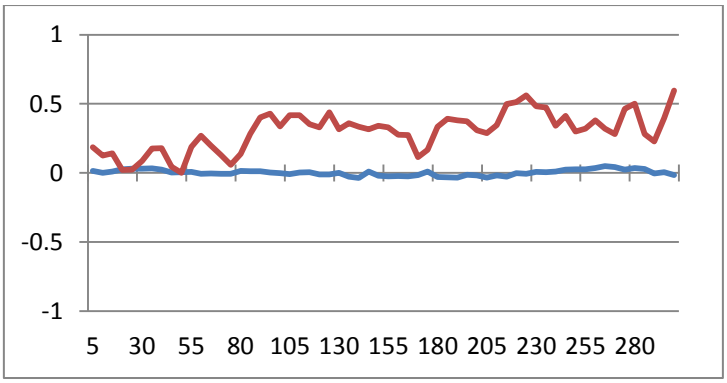

Fig. 54: Correlation of loudness (dB) across time averaged across dyads (x-axis: interaction time in seconds, y-axis: correlation values)

The individual plots in Fig. 53 showed that alignment of loudness varies across time. Speakers sometimes adapt to their partner's loudness, but sometimes they do not. From Fig. 54 it is clear there is also some systematicity within the temporal variation. First, a mixed effects model, with *dyad* as random effect and *real-vs-base* as fixed effect, reveals that there is a global effect of alignment of loudness: the real data differ significantly from the baseline data ($t=18.44$, $p<0.001$). Second, the slight increase in correlation rates we see in Fig. 54 appears to be small, but significant ($t=4.21$, $p<0.001$). This second point was proven by a mixed effects model with *conversation time* as fixed effect.

The results for this study are identical to what we observed for pitch alignment. And again, the results differ from related studies by as Bonin et al. (2013), De Looze et al. (2014) or Kousidis et al. (2009). Although they used a comparable method, they found no overall effect of alignment of loudness, nor were they able to demonstrate an increase over time. This difference in result for alignment of loudness might further be indicative of the relevance of face-to-face contact (as opposed to conversation over the phone) for alignment of prosodic features.

THE TEMPORAL DYNAMICS OF LEXICAL ALIGNMENT

What studies such as Brennan & Clark (1996) or Garrod & Anderson (1987) have demonstrated, is that interlocutors grow routines in referential choices. This means that the more often people refer to a given object, the more likely it is they will use the same word as their conversational partner. In the previous chapter we found converging evidence for this observation, but we also discovered that an increase in alignment over repeated references does not coincide with an increase of alignment over time. In other words, speakers do not align their referential choices more as they talk longer, only as they refer to a same object more often. With the present study we want to further dig into the temporal aspect (not the referential aspect), and broaden the scope from alignment of referential choices to that of lexical alignment in general.

Lexical alignment has been studied both within a corpus linguistics paradigm (Beňuš et al. 2014, Danescu-Niculescu-Mizil et al. 2012, Healey et al. 2014, Manson et al. 2013) and an experimental psycholinguistic

paradigm (Brennan & Clark 1996, Garrod & Anderson 1987, Pickering & Garrod 2004). This paradigm difference not only reflects a difference in research questions and approach, but also a difference in which type of lexical items are being studied: the corpus linguistics studies focus on function words, the psycholinguists study referential choice and thus content words. In this section we will do both.

The method to look into the temporal dynamics of lexical alignment is pretty much identical to what we did for gestural alignment. Our procedure in counting instances of lexical alignment involves three parameters. We consider a prime-target pair to be aligned if prime and target:

(i) consist of the same lemma;
(ii) fall within a 40 second range;
(iii) occur in an interactional pair.

With parameter (i), we want to broaden the scope of lexical alignment beyond a very strict and formal type of alignment. Like we did in Chapter 3 for referential choice, we are measuring alignment of lexical roots, rather than alignment of exact word forms. This means we discard flexion, plural or diminutive markers, conjugation, etc. so as to still match Dutch “mooie” (nice) with “mooi” (nice), “afstandsbediening” (remote control) with “afstandsbedieningen” (remote controls), or “schieten” (to shoot) with “schiet” (shoots). To measure alignment at lexical root level, we used the lemmatised version[46] of our transcription for all of our analyses.

Parameters (ii) and (iii) are exactly the same as in our measuring technique for gestural alignment (cf. supra). We do not want to consider any two identical words that are minutes apart in the transcription as cases of alignment, i.e. as instances in which a speaker adapts his behaviour to that of his partner. Furthermore, if a speaker 1 uses “mooi” (nice) two times, followed by speaker 2 who uses “mooi” (nice), we do not count this

[46] All transcriptions were tagged for parts of speech (POS) by the Frog tagger (Van den Bosch et al. 2007). Part of the output of the tagger was a lemmatised version of the transcriptions. This output was used for the further analyses of lexical alignment.

as two instances of lexical alignment. We only consider two words to be aligned if they are adjacent.

Given the three parameters sketched above, we calculated TAMA plots for the amount of lexical alignment per dyad. Consistent with the rest of this section we used a 40 second *window* and a 5 second *step* for all of the TAMA plots. To avoid calculating an alignment rate based on only a few words, we only considered 40 second windows in which both speakers utter at least 30 words. This kind of data sparseness occurred in 8 per cent of the data points.

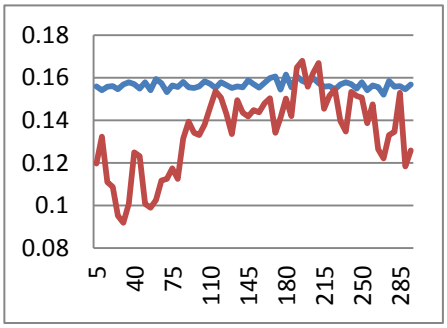

Fig. 55a: TAMA plot for function words

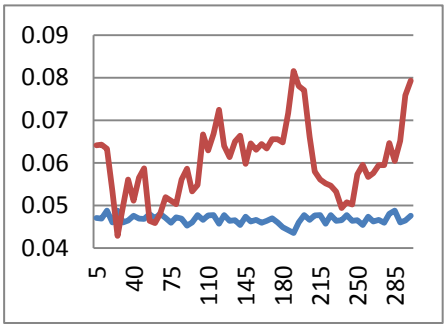

Fig. 55b: TAMA plot for content words

For this study we used POS-tags to differentiate between function and content words. For the latter we only considered adjectives, adverbs, nouns and main verbs; for the former we considered all the rest (i.e. auxiliary verbs, prepositions, pronouns, articles, conjunctions and interjections). When we average the individual TAMA plots across dyads, we observe quite a difference between alignment of function words (Fig. 55a) and alignment of content words (Fig. 55b): content words are aligned systematically more often than in a time-randomised baseline but function words are not. For content words, a mixed effects model with *dyad* as random effect and *real-vs-base* as fixed effect, confirms this difference between real data and baseline is significant ($t=9.98$, $p<0.001$). What we do not see in the averaged TAMA plot in Fig. 52b is a gradual temporal pattern: an increase (or decrease) of alignment of content words across time does

not seem to occur systematically in our data. This is also confirmed by a mixed effects model with *conversation time* as fixed factor (t=1.05, p=0.06).

Within the category of function words, we did not observe any alignment above chance level. The observation of Manson et al. (2013), who found that not all function words reached significant alignment, made us further differentiate within this category. Fig. 56 shows the TAMA plot for all function words except for pronouns and articles. Within this subset of function words, we do observe more alignment than is to be expected by chance (t=2.97, p=0.003). Also, we see a general temporal pattern: at the beginning of the conversation there is no substantial amount of alignment but from about one and a half minutes into the brainstorm there is. A mixed effects model with *conversation time* as fixed factor shows this overall increase of alignment rates is significant (t=8.57, p<0.001). What the difference between Fig. 56 and Fig. 55a illustrates, is the relevance of the *content confound* issue (see also section 3.3.2 in the previous chapter). When uttering pronouns and articles it is very difficult, nearly impossible even, for speakers to not align. For example, speakers who want to refer to themselves using a pronoun are very limited in their possibilities. If a language only offers one way of expressing a piece of content, then can we truly speak of alignment if two speakers in a conversation use that expression? If two speakers use the personal pronoun “I” is that truly an act of adaptive behaviour? Or is the behaviour constrained by the lack of alternatives the linguistic system has to offer? Because pronouns and articles are so frequent, they carried across the content confound issue to the entire category of function words, creating a TAMA plot well below the baseline.

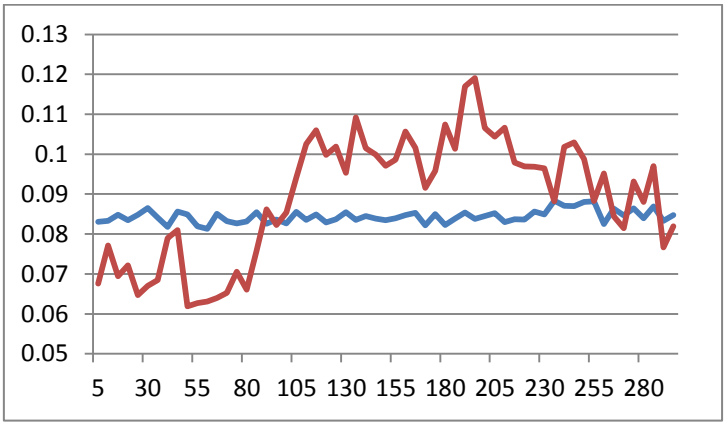

Fig. 56: Averaged TAMA plot for alignment function words, pronouns and articles excluded (x-axis: conversation time in seconds; y-axis: alignment ratio)

Very few studies have systematically looked into the temporal dynamics of lexical alignment. The ones that did either treated the temporal or the linguistic aspect in a very coarse-grained manner. Bonin et al. (2013) for example, did use a TAMA method to capture complex temporal dynamics, but they measure lexical alignment in terms of overlap between adjacent turns. This means, for every turn, they measure the proportion of overlapping words with the previous turn (by the other speaker). Because of this technique, only short range, turn-by-turn alignment is taken into account. Moreover, Bonin and colleagues do not make use of lemmatised transcriptions and therefore measure a very strict and formal type of alignment. For example, a speaker 1 using “vehicles”, followed by speaker 2 using “vehicle” in the next turn, would pass beneath their alignment radar. Also, Bonin et al. (2013) did not differentiate between function and content words. As our results suggest, this is at the risk of frequent function words skewing the alignment rates into an undesirable direction. In the specific case of Bonin and colleagues this risk is even higher because they reduced all personal pronouns to a single form. This means any personal pronoun uttered by speaker 1 will match any other personal pronoun by speaker 2, even further inflating the alignment rates, and feeding the content confound issue.

Some studies, like Manson et al. (2013), are more fine-grained from a linguistics point of view (i.e. they measure alignment rates for different

types of function words separately), but they are very coarse-grained in plotting the observed alignment along a time axis. Manson and colleagues considered three 60 second windows of analysis: one at the beginning, one in the middle and one at the end of the 10 minute conversations they analysed. Given the dynamic character of the alignment rates (at any level, including the lexical) we observe in our study, sampling the data as crudely as they did, might over- or underestimate an effect of increasing alignment over time.

Our results for content words, i.e. interlocutors do align more often than chance but do not align more as the interaction unfolds, tie in with what we observed for referential alignment in Chapter 3. In that chapter we saw that interlocutors typically use the same words to refer to the same objects (viz. the objects in the video animations), but that this type of alignment does not significantly increase over time (neither within individual animation descriptions, nor over the entire experiment). For function words, we did find such an effect of convergence over time in the present study. A potential reason for this difference in outcome might be that repeated references, and by extension all content words, are not evenly distributed across time. They rather occur in clusters that can be linked to conversational topics. If interlocutors are first talking about cars, and then about gardening, this can increase chances of observing two peaks in content word alignment, rather than a gradual increase. Function words however, are more strongly linked to a speaker's individual style (Kestemont 2013) and much less tied to conversational topics. In this vein, very local peaks of stylistic adaptation seem much more unlikely than peaks of referential adaptation.

THE TEMPORAL DYNAMICS OF SYNTACTIC ALIGNMENT

Syntactic alignment has received a substantial amount of attention in the alignment literature (for an overview, see section 1.3.1). Most studies found that speakers indeed copy each other's syntactic constructions. What has not yet been studied is how this syntactic alignment develops over conversation time. Syntactic alignment can be measured in many different ways. Some researchers focussed on alignment of specific syntactic phenomena such as passives, future tense markers, particle placement, etc.

(Gries 2005, Szmrecsanyi 2005). Others used word class n-grams (Dale & Spivey 2006) or syntactic parse trees (Healey, Purver & Howes 2014) to not restrict themselves to a range of specific phenomena, but to study syntactic alignment as a whole. For the present study, we will use the POS-tags in the Insight Interaction Corpus to perform an analysis closely linked to that of Dale & Spivey (2006).

Our method to look into the temporal dynamics of syntactic alignment closely resembles that at the lexical or gestural level (cf. supra). This means we first counted all aligned prime-target pairs that:

(i) consist of the same POS n-gram
(ii) fall within a 40 second range
(iii) occur in an interactional pair

For this study we restricted ourselves to looking at bigrams and trigrams of POS-tags, because of the results in Dale & Spivey (2006): unigrams do not show more alignment compared to a baseline, and for n-grams larger than 3 the difference between alignment in the real data and in the baseline does not further increase if n increases. We first calculated individual TAMA plots (cf. supra) for bigrams and trigrams and then averaged across dyads. Fig. 57a shows the plot for bigrams and Fig. 57b that for trigrams.

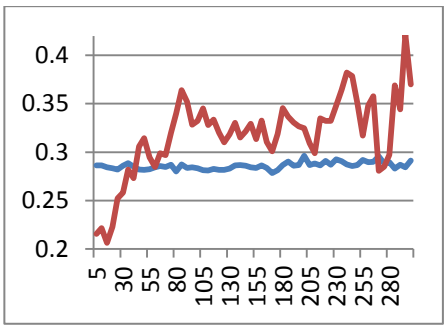

Fig. 57a: TAMA plot for POS bigrams

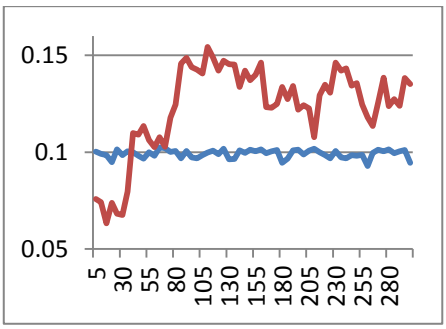

Fig. 57b: TAMA plot for POS trigrams

The plots in Fig. 57 are indicative of a general temporal pattern for syntactic alignment: from the beginning of the brainstorm until about one and a half minute into the conversation, we observe an increase in syntactic alignment. Only after this one and a half minute point in the brainstorm

does the plot for the real data (red line) catch up with that for the time-randomised data (blue line). Both for bigrams and trigrams the alignment rates do not further increase beyond this point, but fluctuate as the conversation continues. Mixed effects models with *dyad* as random effect and *conversation time* as fixed effect, prove the observed increase in alignment rates is significant for both the bigrams ($t=5.05$, $p<0.001$) and the trigrams ($t=5.39$, $p<0.001$). Also, the difference between the alignment rates in the real data and the baseline data is significant for the bigrams ($t=7.59$, $p<0.001$) and the trigrams ($t=12.94$; $p<0.001$).

With this study, we have demonstrated that the speakers in the brainstorm task adapt their syntactic construal to that of their conversational partners more than was to be expected by chance. The comparison to a randomised baseline was not performed in most studies on syntactic alignment (e.g. Bock & Griffin 2000; Branigan, Pickering & Cleland 2000; Gries 2005; Szmrecsanyi 2005). This lack of a baseline comparison was criticised by Healey, Purver & Howes (2014), who claim that such a baseline is indispensable, and who provide corpus-based evidence that syntactic alignment does not occur more often than chance. This study, together with the results in Dale & Spivey (2006), in turn, appears to refute the claims made by Healey and colleagues, and provides further evidence for the presence of syntactic alignment in conversation. Even more, we show how this syntactic alignment is not stable throughout a conversation, but that it first increases above chance level and then fluctuates as the interaction continues.

4.2.3 General discussion on the temporal dynamics of alignment

In this section 4.2 we have demonstrated for a multimodal range of phenomena that alignment is highly dynamic. People do no copy each other's behaviour all the time, instead, alignment rates vary a lot throughout time. For some phenomena we found that alignment increases over time (gesture, pitch, loudness, function words, syntax), for others there was no such global pattern (speech rate, content words).

From a methodological point of view, we have motivated why a fine-grained approach such as the TAMA method can be relevant for research on alignment. Splitting conversations in halves or quarters, or

sampling windows of analysis from different parts of a conversation does not sufficiently do justice to the temporal variability of alignment at different levels. In this fourth case study on the temporal dynamics of alignment we have only shown *that* and *how* alignment rates fluctuate as a function of time, but not *why*. If we are able to reliably plot the amount of alignment along a time axis, then it is also possible to study which other phenomena or factors (that themselves are not stable throughout time) co-vary with it. One way of doing that would be to qualitatively analyse conversations for marked peaks or pits of alignment. Another option would be to create rating experiments in which participants look at video snippets of the recorded conversations. When also applying a TAMA-like method to those ratings, i.e. when calculating average rating scores based on overlapping windows of analysis, a comparison with alignment rates would be easy to make. For example, does conversational dominance or involvement, or perceived naturalness of the conversation fluctuate in parallel with alignment? Or do the parts of a conversation in which people are perceived of being untrue, impolite or humorous correspond to the ones in which a lot of alignment is measured? Temporal approaches such as the one deployed here, open up these interesting new avenues for future research.

A second methodological issue pertains to the relevance of the conversational setting. Our results for prosodic alignment suggest that alignment rates in face-to-face settings differ from that of phone conversations. The robustness of this difference, and whether such a difference might be due to acoustic, social, cognitive or other factors is food for further research. What this case study has exemplified at the very least is that generalising results on alignment from one interaction setting to another or to human communication in general, is risky and should be treated with care.

From a theoretical point of view, the results in this section provide some counter evidence for the rigid and automatic view on alignment processes as in Chartrand & Bargh (1999) or Pickering & Garrod (2004, 2006). If the cognitive mechanism behind adaptive behaviour would only be priming based, alignment rates would not vary a lot throughout conversation time. In the case of repeated references, structural priming

could explain an increase of alignment rates over time, but when making abstraction of this type of referential repetition, a strictly priming based account of communication would not allow much space for different alignment rates in different parts of a conversation. Our results indicate the exact opposite: alignment rates vary radically, i.e. vary between absence of alignment and near perfect alignment, across time. Priming alone cannot account for this type of variation. In this sense, founding a theory on how alignment works on priming alone might not do sufficient justice to the active engagement of interlocutors. As very aptly put by Healey, Purver and Howes (2014: 2) "successful communication seems to depend on the ability to selectively repeat some of our conversational partner's words in different syntactic contexts in order to produce the contrasts, elaborations and corrections that move a conversation forward". Priming might be a relevant but certainly not the only and maybe even not the most important factor underpinning the dynamics in face-to-face conversation.

Chapter 5

Integrating the temporal and multimodal dimension

Throughout this dissertation, we have sporadically coupled temporal and multimodal aspects of alignment in face-to-face interaction. In Chapter 3, for example, we showed how temporal factors such as *distance* between prime and target, and temporal *position* within a conversation or within the experiment, may be predictors at different multimodal levels (i.c. lexical and gestural alignment). In this chapter we want to systematically look into this coupling of temporal and multimodal aspects by comparing the temporal analyses that were done in Chapter 4 across modalities. In doing so, we will unravel how alignment rates at one level correlate with alignment rates at another level. For example, are peaks of lexical alignment time-aligned with peaks at the gestural level? Do speakers who are locally aligning in terms of pitch also more intensely coupled in terms of loudness? Or are there any negative correlations where alignment at one level typically occurs when there is absence of alignment at another level?

5.1 Case study 5: temporal alignment of multimodal alignment

5.1.1 Introduction & research questions

To the best of our knowledge, the only study that systematically focusses on how alignment at different multimodal levels co-varies across conversation time is Bonin et al. (2013). They studied the interplay between lexical and prosodic alignment (at different levels including speech rate, pitch and loudness) and found the two levels under scrutiny to run independently from one another.

Although hardly any attention has been paid to the correlation of alignment rates at different levels, it is a relevant issue from a theoretical point of view. If Pickering and Garrod (2004: 4) claim that "alignment at one level leads to alignment at other levels", then such a link between alignment rates should be apparent from our data. An example of this percolation of alignment throughout different levels could be the lexical boost effect: if speakers copy each other's words, they will also be copying each other's syntactic constructions (or the other way around). With this study we want to broaden the scope of comparisons in a multimodal sense. Words and syntactic constructions can be regarded as dimensions within

the same modality, or even as not being distinct categories at all[47] (as is the case in construction grammar: Goldberg 1995, 2006; Michaelis 2006, Östman & Fried 2005). With the present study we are mainly interested in whether alignment at one level leads to alignment (or the absence thereof) at another *multimodal* level. Of course we are aware of the fact that we cannot look into *causal* relations between alignment at different levels with the set-up in this study, but at least we can provide empirically motivated evidence for the *correlation* between them. In doing so, we can add data-driven evidence for the theoretical claim by Pickering & Garrod.

Because previous research on the topic is very scarce, this chapter will be highly exploratory and not start from a well-defined set of hypotheses. However, on the basis of previous chapters in this work and some related work, we are able to spell out some expectations:

(i) Low correlation between lexical and gestural alignment rates. In Chapter 3, for referential gestures and words, we found that speakers or target objects that are often aligned gesturally are not systematically aligned lexically. We expect a similar pattern in the brainstorm task for content words and depictive gestures.
(ii) Low correlation between lexical and prosodic alignment because Bonin et al. (2013) found these two levels to be unrelated.
(iii) High correlation between lexical and syntactic alignment because of the lexical boost effect (cf. supra).
(iv) High correlation between mutual gaze and lexical alignment because in Chapter 3 we found that eye contact enhances alignment of referential lexical choices.

5.1.2 Method & analysis

In this chapter we will study how alignment at one level correlates with alignment at another. Because this is an exploratory study, we will in fact study how alignment at any level correlates with alignment at any other level. The levels under scrutiny are the ones presented in case study 4 in

[47] This is the case in for example Construction Grammar in which a strict separation between syntax and lexicon is denied. Words and syntactic constructions are seen as part of a *syntax-lexicon continuum*. They are both form-meaning pairs and do not differ structurally but only in terms of internal (symbolic) complexity.

the previous chapter, viz. alignment of depictive gestures, function words, content words, POS n-grams, speech rate, pitch and loudness. Because we expect gaze behaviour to potentially correlate with alignment behaviour[48], we also include one eye gaze feature, viz. eye contact rates.

Before proceeding with the core analyses in this chapter we will first briefly explain how we integrated this factor of eye contact with the rest of the data. Analogous to how we measured the alignment rates at the different multimodal levels in Chapter 4, here we also applied the TAMA method to calculate the amount of eye contact as a function of conversation time. We again used a 40 second *window* of analysis and a 5 second *step* and calculated the relative amount of eye contact (i.e. the total amount of seconds both speakers are looking at each other's face divided by 40 seconds) for every window of analysis and for every dyad. In itself these plots are difficult to interpret; they only indicate how much eye contact there is throughout a conversation, but when coupled to alignment rates, they might show how gaze dynamically affects alignment at different levels.

To check for correlations between alignment at different levels (including eye contact, as just explained above) we will use two techniques. First, we will compute a general correlation measure for every possible multimodal pair. We obtain this pairwise correlation measure by taking the following steps:

(i) The starting point are the alignment rates at the different multimodal levels. Fig. 58 provides an overview of these data for one dyad in the brainstorm task. Plotting all levels under scrutiny would make for an indecipherable graph, so we chose three levels to prove the point. In the example in Fig. 58 we see a clear positive correlation between alignment of pitch and gesture, but negative correlations with eye contact. Note that all alignment rates (and eye-contact) are normalised (by calculating z-scores) to allow for maximal comparability between the different levels.

[48] In the first case study (Chapter 3), but also in Wang et al. (2013) and Postma et al. (2013), we have seen that gaze behaviour indeed does affect lexical and gestural alignment.

(ii) For each possible pairwise combination of multimodal levels, we compute the (Pearson) correlation value. See Table 8 which is based on the data for the dyad in Fig. 58. In table 8 all the possible combinations in Fig. 58 (i.e. pitch_gesture, pitch_eye-contact, gesture_eye-contact) are marked in grey.

(iii) Tables such as the ones in Table 8 are generated for every dyad in the corpus and then averaged to obtain an overall result.

(iv) To test for significance we repeated the procedure above for our time-randomised fictive interactions. This allows us to compare the correlation values in the real data to those for the shuffled data.

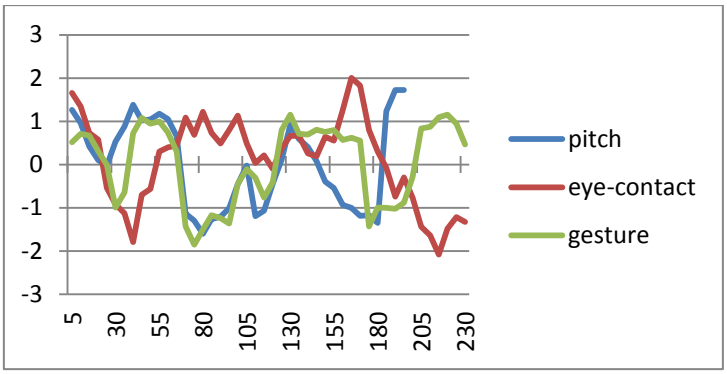

Fig. 58: TAMA plots of alignment at different levels for one dyad in the brainstorm task

	speech rate	pitch	loud ness	fct word	cont word	POS	eye cont	gest
speechrate	*1.00*							
pitch	-0.76	*1.00*						
loudness	-0.05	0.00	*1.00*					
fct_word	0.60	-0.27	-0.09	*1.00*				
cont_word	0.20	-0.32	-0.10	0.05	*1.00*			
POS	0.68	-0.35	0.22	0.76	0.19	*1.00*		
eye cont	0.45	-0.50	-0.36	0.11	0.35	0.38	*1.00*	
gesture	-0.39	0.48	-0.08	0.11	-0.15	0.19	-0.23	*1.00*

Table 8: All pairwise correlation coefficients (Pearson) for the dyad in Fig. 58

The procedure described above provides a global view on the pairwise correlations between alignment rates. However, this analysis does not fully take into account the temporal dynamics of alignment. Just as the alignment rates themselves, maybe also the correlations between those alignment rates occur in local peaks. For example, the correlation between lexical and gestural alignment across time might be low overall, but high during some parts of the conversation. To check for this type of correlation we performed a cluster analysis by taking the following steps:

(i) The starting point of this cluster analysis is the same as the one in the correlation analysis above, viz. the normalised alignment rates (supplemented with the eye contact analysis) that resulted from the TAMA analyses in Chapter 4 (cf. supra, see also Fig. 58).

(ii) For each data point, i.e. for every 5 second step, we check whether or not there is a markedly large or small amount of alignment. We do that by highlighting all values lower than -1 or higher than 1 for each of the levels under scrutiny[49]. Fig. 59 visualises how this works: the plot is based on the data of one dyad (i.c. on the same as in Fig. 58) and shows where during the interaction there are markedly high (value higher than 1) and markedly low alignment rates (values lower than -1) for the three levels shown in this example.

(iii) For all of the possible two-way, three-way and four-way combinations of our eight levels (speechrate, pitch, loudness, content words, function words, POS-tags, eye contact, gesture) we then compute how often each combination occurs with all the items in the combination having high (>1) or low (<-1) scores. For the two-way comparisons we also computed combinations in which one level has a high and the other has a low score. Most obvious from the example in Fig. 59 is that high values for gesture co-occur with high values for pitch, low gesture values with low pitch values, and low pitch values with high eye contact values. This is consistent with the results from the correlation analysis (see Table 8). The

[49] Because we are working with z-scores, the values of 1 and -1 make intuitive sense: z-scores higher than 1 represent data points that are more than 1 standard deviation higher than the mean, i.e. they are markedly high. In the entire data set roughly 33% of the data points are higher than 1 or lower than -1.

difference resides in the approach, which enables us to capture local clusters of alignment for levels that are not or only moderately correlated as a whole.

(iv) When averaging across dyads, the procedure allows us to check which of all the possible clusters of high, low or opposed alignment rates occur frequently in our data.

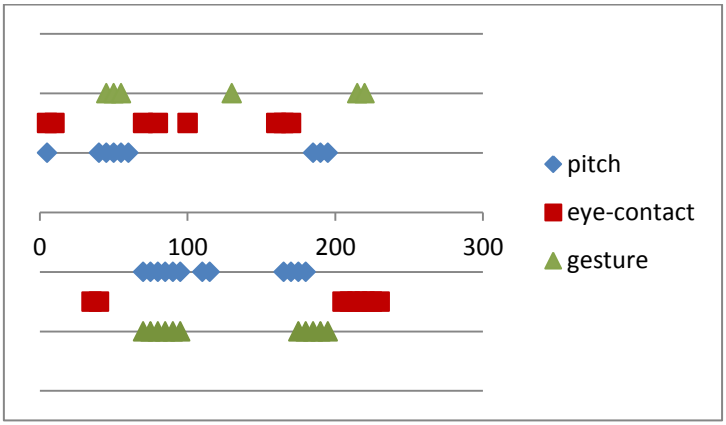

Fig. 59: Plot of alignment peaks and dips for the data in Fig. 58 (above the x-axis (time in seconds) are data points >1; below data points <-1)

5.1.3 Results

In Table 9 we show an overview of our correlation analysis. The table shows the correlation values for each possible two-way combination of alignment levels, averaged across dyads. Marked in grey are the correlation values that differ significantly (t-test with $p<0.01$) from the values in the time-randomised baseline data. Only four combinations appear to correlate systematically: pitch and loudness, pitch and POS n-grams, POS n-grams and function words, POS n-grams and loudness.

	speech rate	pitch	loud ness	fct word	cont word	POS	eye cont	gest
speechrate	*1.00*							
pitch	-0.09	*1.00*						
loudness	0.00	0.22	*1.00*					
fct_word	-0.04	0.11	0.06	*1.00*				
cont_word	-0.02	0.10	-0.12	0.14	*1.00*			
POS	-0.02	0.22	0.26	0.37	0.08	*1.00*		
eye cont	0.08	0.02	0.07	0.06	-0.12	0.16	*1.00*	
gesture	-0.03	0.03	0.06	-0.05	0.06	0.03	0.02	*1.00*

Table 9: (Pearson) correlation measures of alignment rates for all possible pairs

For the results of our cluster analysis, we will only present data for the two-way combinations. For the three- and four-way combinations we hardly ever measured alignment rates higher than 1 or lower than -1. For example, the largest effect we found was for the combination loudness-pitch-POS. However, out of the 453 loudness-pitch-POS combinations in the data, in only 9 cases all three levels had alignment rates with a value higher than 1. For all other three- and four-way combinations we found even less clusters of alignment rates higher than 1 or lower than -1. Ranking the combinations according to their alignment cluster frequency, then, would not make sense for these three- and four-way combinations.

The two-way combinations clustered sufficiently often to make such a ranking useful. Table 10 shows the five most frequently occurring clusters of markedly high alignment rates (both values higher than 1), markedly low alignment rates (both values lower than -1) and markedly opposing rates (one value higher than 1 and the other lower than -1). However, and maybe most importantly, for none of the two-way combinations in Table 10 do we measure significantly more alignment clusters than in our randomised baseline.

	cluster high	cluster low	cluster opposed
1	loudness-POS	eye contact-speechrate	pitch-speechrate
2	loudness-pitch	fct_word-POS	cont_word-speechrate
3	fct_word-POS	loudness-pitch	fct_word-speechrate
4	cont_word-fct_word	fct_word-pitch	gesture-fct_word
5	POS-speechrate	pitch-POS	POS-speechrate

Table 10: Top five most frequent alignment clusters

5.1.4 Discussion

Especially for lexical alignment we had some hypotheses concerning the clustering with other alignment levels. First, in line with Bonin et al. (2013), we did not expect lexical alignment to correlate with alignment of prosodic features. This is confirmed by both the correlation and the cluster analysis. The only link between the two levels we found, was that function words and pitch show some overlap in the negative extremes (see Table 10): low alignment rates at the one level co-occur with low levels at the other. However, this effect is only a tendency since it failed to reach statistical significance when compared to the baseline data.

Second, and drawing on the results in Chapter 3, we expected a low correlation between lexical and gestural alignment. This hypothesis is confirmed by the present study. As is apparent from Table 9, the two levels are not correlated at all. If anything, there is an opposite relation between them (see Table 10): high alignment rates (>1) at one level frequently co-occur with low alignment rates (<-1) at the other level. This effect, however, is just a trend and did not reach statistical significance.

Third, we found a link between lexical and syntactic alignment: in Table 9 we see a significant correlation between alignment of function words and of POS n-grams. This correlation is also reflected in Table 10, where we observe that both peaks and pits of alignment rates at the two levels co-occur frequently. Bock (1986) found syntactic alignment to occur regardless of the lexical content of prime and target. Others, such as Branigan, Pickering & Cleland (2000) or Cleland & Pickering (2003), found a *lexical boost effect*: syntactic alignment was more likely to occur when prime and target contain the same (or conceptually related) lexical items. The present study also provides arguments against the independency of

lexical and syntactic alignment claimed by Bock (1986). However, Pickering and colleagues focussed on the role of content words, and more specifically on the verbs and nouns in a limited set of verb- and noun-tied constructions. With the present study we provide a piece of evidence that the lexical boost effect also applies to function words and is not restricted to a rather small set of syntactic constructions.

Fourth, because of what we found in Chapter 3, we expected a link between lexical alignment and eye contact. However, neither the correlation nor the cluster analysis confirmed this expectation. The absence of an effect in this case study, proves the relevance of our more fine-grained analysis in Chapter 3. What we showed there was that not just eye contact, but more specifically the gaze behaviour of the speaker during the prime, was a good predictor for lexical alignment. Eye gaze in this case study was probably measured too coarsely (viz. the average amount of mutual gaze per 40 second window) to observe a comparable effect. Apparently, in the interplay between gaze and lexical alignment, the correlation resides in individual gaze events that are closely time-aligned rather than in average amounts of mutual gaze.

An important observation from the results in Table 9 (and 10) is that the significant effects we find all result from intra-modal comparisons. Only alignment rates within the speech modality are significantly correlated. Alignment rates between the modalities of gaze, speech and gesture do not appear to be linked. If Pickering & Garrod (2004:7) claim that “interlocutors will tend to align expressions at many different levels at the same time”, this claim does not appear to stretch to a multimodal level. The conversational phases in which there is a lot of gestural alignment do not systematically correlate with phases of high syntactic alignment rates or high eye contact rates. Pickering and Garrod do not include modalities beyond the speech signal in their interactive alignment model, but even among the different levels of speech-tied alignment, we do not observe a lot of correlation. If the percolation of alignment from one level to many other levels would be strong, we would have expected to observe higher correlation measures (in Table 9) than we do. Also, we would have expected more complex clusters of alignment peaks and pits to arise from the data. Our results suggest that interlocutors are hardly ever

markedly aligning (or disaligning) at more than two levels (of the ones under scrutiny in this study) at the same time. Within the two-way alignment clusters, we did find some peaks and pits co-occurring, but they failed to reach statistical significance. In sum, what this study suggests is that alignment at one level *sometimes* (but not very often) leads to alignment at *some* other levels (but not to many), and only if those levels pertain to the same multimodal layer. These results at the least nuance the claim by Pickering & Garrod that there is a link between alignment at different levels, if not they provide empirically motivated counter-evidence for it.

Conclusion

In this dissertation we have set out to explore the multimodal and temporal dimension of alignment. The fact that interlocutors copy each other's behaviour at many different levels had already been established. How alignment at one level is affected by (alignment) behaviour at another level only received little attention in the literature so far. With this dissertation we put some first steps into this direction. Overall, we found very little links between the levels of analysis under scrutiny. Especially across the main modes of representation, viz. speech, gesture and gaze, hardly any relation was found:

(i) speakers that align often gesturally do not systematically align often lexically;
(ii) target objects that are often aligned gesturally are not systematically aligned often lexically;
(iii) lexical and gestural alignment are predicted by different factors;
(iv) gaze affects lexical and gestural alignment in a totally different way;
(v) alignment rates fluctuate over conversation time. However, no correlations between gestural alignment or speech alignment could be observed.

Together, these observations provide converging evidence that alignment is a multifaceted phenomenon that is not governed by a single factor or by a fixed set of factors. Rather, different factors explain alignment at different levels. As illustrated in the final chapter, alignment is not only multifaceted but also very dynamic. Even within the different levels under scrutiny, we measured a lot of variation in alignment rates across conversation time. Between those levels, only little correlation appeared to exist and also our cluster analysis demonstrated that the local peaks and pits of alignment do not bunch up more often than they would do by chance. These insights give rise to some crucial objections against all too rigid accounts of the mechanisms underpinning alignment. First, strictly priming based models of alignment such as the one in Pickering and Garrod (2004, 2006) imply a percolation model in which alignment at one level leads to alignment at other levels. At the least, our findings limit this notion of percolation to

monomodal proportions: the observations in (i)-(v) indicate that a cross-modal percolation of alignment (i.e. across gaze, gesture and speech) does not exist. Second, if priming were the only factor shaping interactive alignment, we would expect it to decrease over distance. Based on the results in Chapter 3, we can nuance this: with the exception of co-constructed gestures (in which the distance between prime and target is 0), we did not find *distance* to affect alignment. Although we admit it is difficult to interpret non-significant results, it does add to the assumption that priming alone does not suffice to explain what we observe in our data. We argue that speakers not (only) align because they were primed, but (also) for different reasons: to trump, to persuade, to counter-argue, to mislead, to acknowledge, to impress, to clarify, to question, etc. In building models of alignment, and by extension models of interaction, that not only take cognitive processes into account, but also display attention for more semantic-discursive factors, we might find an interesting challenge for future research. In this sense, this dissertation might be an attempt to at least show the relevance that more bridges between cognitive accounts and linguistic or communicative accounts are needed to fully grasp the phenomenon.

From a methodological point of view, we have contributed to the research on alignment in at least the following four ways. First, we have tried to take some of the observations made in controlled lab settings *into the wild*. For example, because we used head-mounted eye-trackers, we were able to combine a face-to-face conversational setting with very accurate eye gaze measurements. This allowed us to nuance some of the findings from experimental studies such as the interaction between gaze and gestural alignment: the observation of Wang et al. (2011, 2014) that participants who are being watched by an actor in a video are faster at performing an aligned gesture was contradicted in our face-to-face setting in which AddresseeGaze during the prime did not affect gestural alignment. Another merit of the mobile eye-tracking technique was that we were able to observe the so-called *gaze cueing* effect: speakers fixate their partner's gesture because that partner has just fixated his own gesture. Although not a shocking observation, this had not yet been empirically validated in actual face-to-face settings.

Second, throughout the dissertation we took care not to commit type I errors, i.e. we tried avoiding to detect effects that were actually not there, by structurally comparing our results to baseline results in a fictive data set of time-randomised conversations. This technique proved to be useful and sometimes even necessary. For example, to compare cross-recurrence rates in one data set (e.g. the brainstorm task) to those in another data set (e.g. the animation description task) it did not suffice to simply compare the raw recurrence rates. Because those rates are dependent on the frequency of the phenomenon, it was necessary to calculate an alternative measure to reliably compare results across data sets. This alternative measure was obtained by computing the difference between the recurrence rate in the real data and in the baseline data for each data point.

Third, we further developed the TAMA-method proposed by Kousidis et al. (2008) to fit our research needs. The fine-grained investigation of the temporal dynamics of alignment allowed us to answer the two main questions on the temporal dimension: does alignment increase over time, or does it occur in local peaks? It appears to do both. For all levels we found that alignment rates are not stable but develop dynamically over conversation time. In some parts of the conversations there is a lot of or even full alignment, in other parts there is only little or even no alignment at all. Besides the observation of these local peaks and pits, for some levels we did find a systematic temporal pattern: when averaged across dyads, gestural, lexical and syntactic alignment significantly increase over conversation time. This increase, however, is gradual for lexical alignment, but rather sudden (about one and a half minute into the brainstorm) for lexical and gestural alignment. The temporal aspect of alignment has been largely neglected in the literature so far, and TAMA-like approaches might be a gateway into this issue.

A final result with methodological implications is the observation that different tasks or conversational settings can yield different results. In the temporal coupling between interlocutors' eye gaze, we found a significant difference between the animation description task and the brainstorm task. Also, some of the results obtained in our case studies might differ from the results in related studies due to a difference in conversational setting. The difference between face-to-face interaction and

computer (screen) mediated interaction might influence the amount, direction or temporal dynamics of alignment. Because psycholinguistic research into alignment and synchronisation typically uses task-based interaction types such as the one in our animation description task, we have to critically question to what extent the results can be generalised from the particular data set to human communication in general. We therefore argue that research on synchronisation and alignment can benefit from studying the phenomenon in more diverse interactional settings. This is a further plea to test the robustness of results obtained in experimentally controlled conditions in the messiness of more spontaneous, or at least a range of different interaction types.

In the chapters 3-5 we answered most of the research questions we put forward. However, those answers in turn lead to new questions. For example, although our TAMA approach provided some useful insights, it might also enable us to build bridges between more quantitative and qualitative analyses of alignment: if we quantitatively observe local peaks of (clustered) alignment, can this be related to a qualitative discourse analysis? Or could we link the temporal analysis of alignment with temporal analyses or ratings of other phenomena? For example, do conversational dominance, perceived liking, or even heart rate measurements co-vary with alignment over interaction time? This is interesting from a theoretical/analytical point of view, but also from a methodological perspective because it forces us to further define what we count as alignment and how we link this counting to the time axis of the ongoing conversation. Another obvious next step is to further dig into gestural alignment. Throughout this dissertation we treated gesture very holistically and there is much more to be learned about which gesture features are susceptible to alignment, which features co-vary with alignment at other levels or how gestural alignment is relevant in face-to-face interaction altogether. A final line of following up on the research in this dissertation would be to further test the results in different conversational settings. In the Insight Interaction Corpus the interlocutors are peers that know each other really well who are engaged in a cooperative task. What impact would a different relation between the interlocutors, a non-cooperative or the absence of a task, multiparty interaction, etc. have on how multimodal

alignment shapes interaction? Or how can social, gender, cognitive, emotional, or other information on the interlocutors function as a predictor for who will align to who and to what extent?

From this dissertation it is clear that alignment is a dynamic and multifaceted rather than a static and mechanistic phenomenon in face-to-face interaction. Research on this topic should maximally take that into account. My dad is true chameleon. But he is not pulling the chameleon trick all the time, nor is he constantly doing it at all of the multimodal levels to his disposal. Figuring out which conversational, cognitive and social factors shape this dynamic alignment will remain an interesting challenge for the years to come.

References

Allwood, J. (2008). Multimodal corpora. In: A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook*, 207-225, Berlin: Mouton de Gruyter.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C. & Paggio, P. (2007). The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management, and Sequencing Phenomena. In: J. Martin, P. Paggio, P. Kuenlein, R. Stiefelhagen & F. Pianesi (Eds.), *Multimodal Corpora for Modelling Human Multimodal Behaviour*, 273-287, Heidelberg: Springer.

Al Moubayed, S., Edlund, J. & Gustafson, J. (2013). Analysis of gaze and speech patterns in three-party quiz game interaction. *Proceedings of 14th Annual Conference of the International Speech Communication Association*, 1125-1129.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. &Weinert, R. (1991). The HCRC Map Task corpus. *Language and Speech*, 34, 351-366.

Argyle, M. & Cook, M. (1976). *Gaze and Eye contact*. London: Cambridge University Press.

Baayen, R. (2008). *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Babel, M. (2009). *Phonetic and social selectivity in speech accommodation* (Unpublished doctoral dissertation). University of California, Berkeley.

Bailenson, J. & Yee, N. (2005). Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological Science* 16, 814–819.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-6. http://CRAN.R-project.org/package=lme4.

Bateson, M., Nettle, D. & Roberts, G. (2006). Cues of being watched enhance cooperation in real-world setting. *Biology Letters* 2, 412-414.

Bavelas, J., Coates, L. & Johnson, T. (2002) Listener responses as a collaborative process: The role of gaze. *Journal of Communication* 52, 566-580.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M., Croft, W., Ellis, N., Holland, J., Ke, J., Larsen-Freeman, D. & Schoenemann, T. (2009). Language Is a Complex Adaptive System: Position Paper. *Language Learning* 59, 1-26.

Beňuš, S., Gravano, A., Levitan, R., Levitan, S., Willson, L. & Hirschberg, J. (2014). Entrainment, dominance and alliance in supreme court hearings. *Knowledge-Based Systems* 71, 3-14.

Bergmann, K. & Kopp, S. (2012). Gestural alignment in natural dialogue. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 1326 - 1331.

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B. & Rauzy, S. (2008). Le CID - Corpus of Interactional Data – Annotation et exploitation multimodale de parole conversationnelle. *Traitement automatique des langues* 49, 105-134.

Bock, J. (1986). Syntactic persistence in language production. *Cognitive Psychology* 18, 355-387.

Bock, K. & Griffin, Z. (2000). The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology* 129, 177-192.

Boersma, P. & Weenink, D. (2009). PRAAT: doing phonetics by computer (Version 5.3.05). http://www.praat.org/. Accessed 27 February 2012.

Bonin, F., De Looze, C., Ghosh, S., Gilmartin, E., Vogel, C., Polychroniou, A., Salamin, H., Vinciarelli, A. & Campbell, N. (2013). Investigating fine temporal dynamics of prosodic and lexical accommodation. *Proceedings of Interspeech,* 539-543.

Bourhis, R. & Giles, H. (1977). The Language of Intergroup Distinctiveness. In: H. Giles (Ed.), *Language, Ethnicity and Intergroup Relations*, 119-135, London: Academic Press.

Branigan, H., Pickering, M., McLean, J. & Cleland, A. (2007). Participant role and syntactic alignment in dialogue. *Cognition* 104, 163-197.

Branigan, H., Pickering, M., Pearson, J., McLean, J., Nass, C. & Hu, J. (2004). Beliefs about mental states in lexical and syntactic alignment: Evidence

from human–computer dialogs. Paper Presented at the 17th Annual CUNY Human Sentence Processing Conference.

Branigan, H., Pickering, M. & Cleland, A. (2000). Syntactic co-ordination in dialogue. *Cognition* 75, 13-25.

Brennan S., Chen X., Dickinson C., Neider M. & Zelinsky G. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition* 106, 1465-1477.

Brennan, S. & Clark, H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22, 1482–93.

Brennan, S. & Hanna, J. (2009). Partner-specific adaptation in dialog. *Topics in Cognitive Science* 1, 274-291.

Brewer, M. (1991). The social self: on being the same and different at the same time. *Personality and Social Psychology Bulletin* 17, 475–482.

Brône, G., Feyaerts, K. & Oben, B. (2013). Multimodal turn-taking in dialogue: on the interplay of eye gaze, speech and gesture. *Proceedings of AFLiCo5: Empirical approaches to multi-modality and to language variation*, 21-22.

Brône, G. & Oben, B. (2015). InSight Interaction. A multimodal and multifocal dialogue corpus. *Language Resources and Evaluation* 49, 195-214.

Brône, G. & Oben, B. (2013). Resonating humour: A corpus-based approach to creative parallelism in discourse. In: K. Feyaerts, T. Veale, C. Forceville (Eds.), *Creativity and the agile mind: A multi-disciplinary study of a multi-faceted phenomenon*, 181-204, Berlin: De Gruyter.

Brône, G. & Zima, E. (2014). Towards a dialogic construction grammar. Ad hoc routines and resonance activation. *Cognitive Linguistics* 25, 457-495.

Campana, E., Silverman, L., Tanenhaus, M., Bennetto, L. & Packard, S. (2005). Real-time integration of gesture and speech during reference resolution. *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, 378-383.

Campbell, N. (2008). Tools and Resources for Visualising Conversational Speech Interaction. *Proceedings of the 6th International Conference on Language Resources and Evaluation*, 231-234.

Cassell, J., Torres, O. & Prevost, S. (1999) *Turn Taking vs. Discourse Structure: How Best to Model Multimodal Conversation*. The Hague: Kluwer.

Cassell, J., Sullivan, J., Prevost, S. & Churchill, E. (2000). *Embodied Conversational Agents*, Massachusetts: MIT Press.

Catmur, C., Gillmeister, H., Bird, G., Liepelt, R., Brass, M. & Heyes, C. (2008). Through the looking glass: Counter-mirror activation following incompatible sensorimotor learning. *European Journal of Neuroscience* 28, 1208-1215.

Cavicchio, F. & Poesio, M. (2009). Multimodal Corpora Annotation: Validation Methods to Assess Coding Scheme Reliability. *Lecture Notes in Computer Science 5509,* 109-121.

Chartrand, T. & Bargh, J. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology* 76, 893–910.

Chartrand, T., Maddux, W. & Lakin, J. (2005). Beyond the perception-behavior link: The ubiquitous utility and motivational moderators of nonconscious mimicry. In: R. Hassin, J. Uleman & J. Bargh (Eds.), *The new unconscious*, 334-362, New York: Oxford University Press.

Chen, L., Travis-Rose, R., Parrill, F., Han, X., Tu, J., Huang, Z., Harper, M., Quek, K., McNeill, D., Tuttle, R. & Huang, T. (2006). VACE Multimodal Meeting Corpus. *Lecture Notes in Computer Science* 3869, 40-51.

Clark, H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Cleland, A. & Pickering, M. (2003). The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language* 49, 214–230.

Coco, M. & Dale, R. (2014). Cross-recurrence quantification analysis of behavioral streams: methodology and application. *Frontiers in Quantitative Psychology and Measurement* 5, 1-14.

Collins, B. (1998). Convergence of fundamental frequencies in conversation: if it happens, does it matter? *Fifth International Conference on Spoken Language Processing*.

Costa, A., Pickering, M. & Sorace, A. (2008). Alignment in second language dialogue. *Language and Cognitive Processes* 23, 528-556.

Cummins, F. (2011). Gaze and blinking in dyadic conversation: A study in coordinated behaviour among individuals. *Language and Cognitive Processes* 27, 1525-1549.

Dale, R., Kirkham, N. & Richardson, D. (2011). How two people become a tangram recognition system. *Proceedings of the European Conference on Computer-Supported Cooperative Work*. Berlin: Springer Verlag.

Dale, R. & Spivey, M. (2006). Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning* 56, 391-430.

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B. & Kleinberg, J. (2012). Echoes of power: Language effects and power differences in social interaction. Proceedings of WWW, 699-708.

Danescu-Niculescu-Mizil, C., Gamon, M. & Dumais, S. (2011). Mark my words! Linguistic style accommodation in social media. *Proceedings of WWW,* 745-754.

Darwin, C. (2009). The expression of the emotions in man and animals. In: F. Darwin (Ed.), Cambridge: Cambridge University Press.

De Looze, C., Scherer, S., Vaughan, B. & Campbell, N. (2014). Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. *Speech Communication* 58, 11-34.

Dijksterhuis, A. & Bargh, J. (2001). The perception-behavior expressway: Automatic effects of social perception on social behavior. *Advances in Experimental Social Psychology 33*, 1–40.

Dominey, P. (2004). Situation alignment and routinization in language acquisition. *Behavioral and Brain Sciences* 27, 195.

Du Bois, J. (2010). *Towards a dialogic syntax*. Unpublished manuscript.

Du Bois, J. (2014). Towards a dialogic syntax. *Cognitive Linguistics* 25, 359-410.

Duchowsky, A. (2007). *Eye Tracking Methodology. Theory and Practice*. Berlin: Springer.

Emery, N. (2000). The eyes have it: The neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews* 24, 581–604.

Fowler C., Brown J., Sabadini L. & Weihing J. (2003). Rapid access to speech gestures in perception: evidence from choice and simple response time tasks. *Journal of Memory and Language* 49, 396-413.

Frischen, A., Bayliss, A., & Tipper, S. (2007). Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological Bulletin* 133, 694-724.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C. & Tylén, K. (2012). Coming To Terms: Quantifying the Benefits of Linguistic Coordination. *Psychological Science* 23, 931 - 939.

Fusaroli, R., Konvalinka, I. & Wallot, S. (2014). Analyzing social interactions: The promises and challenges of using cross recurrence quantification analysis. *Springer Proceedings in Mathematics and Statistics* 103, 137-155.

Fusaroli, R., Raczaszek-Leonardi, J. & Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas in Psychology* 32, 147-157.

Garrod, S. & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition* 27, 181–218.

Garrod, S. & Pickering, M. (2004). Why conversation is so easy? *Trends in Cognitive Sciences* 8, 8-11.

Gerwing, J., Allison, M. (2009). The relationship between verbal and gestural contributions in conversation: A comparison of three methods. *Gesture* 9, 312-336.

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In: H. Giles, J. Coupland, & N. Coupland (Eds.), *Contexts of Accommodation,* 1-68, Cambridge: Cambridge University Press.

Giles, H. & Powesland, P. (1975). *Speech styles and social evaluation*. New York: Academic Press.

Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago/London: University of Chicago Press.

Goldberg, A. (2006). Constructions at work. The nature of generalization in language. Oxford: Oxford University Press.

Goldinger, S. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review* 105, 251-279.

Goudbeek, M. & Krahmer, E. (2012). Alignment in interactive reference production: Content planning, modifier ordering and referential overspecification. *Topics in Cognitive Science* 4, 269-289.

Gregory, S. & Hoyt, B. (1982). Conversation partner mutual adaptation as demonstrated by Fourier series analysis. *Journal of Psychological Research* 11, 35-46.

Gries, S. (2005). Syntactic Priming: A Corpus-based Approach. *Journal of Psycholinguistic Research* 34, 365-399.

Gullberg, M. & Holmqvist, K. (2006). What speakers do and what listeners look at. Visual attention to gestures in human interaction live and on video. *Pragmatics and Cognition* 14, 53-82.

Gullberg, M. & Kita, S. (2009). Attention to speech-accompanying gestures: Eye movements and information uptake. *Journal of Nonverbal Behaviour* 33, 251-277.

Hadar, U. (2013). Coverbal gestures: Between communication and speech production. In: C. Müller et al. (Eds.), *Body - Language - Communication. An International Handbook on Multimodality in Human Interaction*, 804-820, Berlin: De Gruyter.

Hadelich, K. & Crocker, M. (2006). Gaze alignment of interlocutors in conversational dialogues. *Proceedings of the 2006 Symposium on Eye Tracking Research and Applications*, 38.

Healey, P., Purver, M. & Howes, C. (2014). Divergence in dialogue. PLoS ONE 9, 1-6.

Herrera, D., Novick, D., Jan, D. & Traum, D. (2010). The UTEP-ICT Cross-Cultural Multiparty Multimodal Dialog Corpus. *Proceedings of the 7th International Conference on Language Resources and Evaluation*, 49-54.

Heyes, C., Bird, G., Johnson, H. & Haggard, P. (2005). Experience modulates automatic imitation. *Cognitive Brain Research* 22, 233-240.

Holler, J., & Kendrick, K. (2015). Unaddressed participants' gaze in multi-person interaction: Optimizing recipiency. *Frontiers in Psychology* 6, 1-14.

Holler, J. & Wilkin, K. (2011). Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior* 35, 133–153.

Hove, M. & Risen, J. (2009). It's all in the timing: Interpersonal synchrony increases affiliation. *Social Cognition* 27, 949-961.

Howes, C., Healey, P. & Purver, M. (2010). Tracking Lexical and Syntactic Alignment in Conversation. *Proceedings of the Twenty-fifth Annual Conference of the Cognitive Science Society*, 2004-2009.

Iacoboni, M., Woods, R., Brass, M., Bekkering, H., Mazziotta, J. & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science* 286, 2526–2528.

Ireland, M. & Pennebaker, J. (2010). Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of Personality and Social Psychology* 99, 549–571.

Jacob, R. & Karn, K. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In: R. Radach, J. Hyönä & H. Deubel (Eds.), *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, 573-605, Oxford: Elsevier Science.

Jakobson, R. (1960). Linguistics and Poetics. In: T. Sebeok (Ed.), *Style in Language*, 350-377, Cambridge MA: M.I.T. Press.

James, W. (1878). Brute and human intellect. *The Journal of Speculative Philosphy* 12, 236-276.

Jokinen, K. (2010). Non-verbal signals for turn-taking and feedback. *Proceedings of the 7th International Conference on Language Resources and Evaluation,* 2961-2967.

Jones, B., DeBruine, L., Little, A., Conway, C. & Feinberg, D. (2006). Integrating gaze direction and expression in preferences for attractive faces. *Psychological Science* 17, 588-591.

Juslin, P. & Scherer, K. (2005). Vocal expression of affect. In: J. Harrigan, R. Rosenthal & K. Scherer (Eds.), *The New Handbook of Methods in Nonverbal Behavior Research*, 65-135, Oxford: Oxford University Press.

Kendon, A. (1967) Some functions of gaze-direction in social interaction. *Acta Psychologica* 26, 22-63.

Kendon, A. (1988). How gestures can become like words. In: F. Poyatos (Ed.), *Crosscultural Perspectives in Nonverbal Communication*, 131-141, Toronto: C. J. Hogrefe Publishers.

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Kestemont, M. (2013). *Het gewicht van de auteur: stylometrische auteursherkenning in Middelnederlandse literatuur*. Gent: Koninklijke Academie voor Nederlandse Taal- en Letterkunde.

Kimbara, I. (2006). On gestural mimicry. *Gesture* 6, 39-61.

Kimbara, I. (2008). Gesture Form Convergence in Joint Description. *Journal of Nonverbal Behavior* 32, 123–131.

Kleinke, C. (1986). Gaze and eye contact: A research review. *Psychological Bulletin* 100, 78-100.

Knight, D. (2011). The future of multimodal corpora. *Revista Brasileira de Linguistica Aplicada* 11, 391-415.

Knight, D., Adolphs, S., Tennent, P. & Carter, R. (2008). The Nottingham multi-modal corpus: a demonstration. *Proceedings of the 6th International Conference on Language Resources and Evaluation*, 26-31.

Konvalinka, I., Xygalatas, D., Bulbulia, J., Schjødt, U., Jegindø, E., Wallot, S., Van Orden, G. & Roepstorff, A. (2011). Synchronized arousal between performers and related spectators in a fire-walking ritual. *Proceedings of the National Academy of Sciences of the United States of America* 108, 8514-8519.

Kopp, S., Bergmann, K. & Wachsmuth, I. (2008). Multimodal communication from multimodal thinking. Towards an integrated model of speech and gesture production. *International journal of semantic computing*, 2, 115-136.

Kousidis, S., Dorran, D., Mcdonnell, C. & Coyle, E. (2009). Convergence in human dialogues time series analysis of acoustic feature. *Proceedings of SPECOM 2009*.

Lachat, F., Conty, L., Hugueville, L. & George, N. (2012). Gaze cueing effect in a face-to-face situation. *Journal of Nonverbal Behaviour* 36, 177-190.

Langton, S., Watt, R. & Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in cognitive sciences* 4, 50-58.

Lakin, J., Chartrand, T. & Arkin, R. (2008). I am too just like you: Nonconscious mimicry as an automatic behavioral response to social exclusion. *Psychological Science 19*, 816-822.

Lakin, J., Jefferis, V., Cheng, C. & Chartrand, T. (2003). The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of Nonverbal Behavior* 27, 145-162.

Lausberg, H. & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods, Instruments, & Computers* 41, 841-849.

Levelt, W. & Kelter, S. (1982). Surface form and memory in question answering. *Cognitive Psychology* 14, 78-106.

Lewandowski, N. (2012). *Talent in nonnative phonetic convergence* (Unpublished doctoral dissertation). Universität Stuttgart, Stuttgart.

Louwerse, M., Dale, R., Bard, E. & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive Science* 36, 1404-1426.

Lücking, A., Bergmann, K., Hahn, F., Kopp, S. & Rieser, H. (2013). Data-based analysis of speech and gesture: the Bielefeld Speech and Gesture Alignment corpus (SaGA) and its applications. *Journal on Multimodal Interfaces* 7, 5-18.

Manson, J., Bryant, G., Gervais, M. & Kline, M. (2013). Convergence of speech rate in conversation predicts cooperation. *Evolution and Human Behavior* 34, 419-426.

Marcus, M., Kim, G., Marcinkiewicz, M., MacIntyre, R., Bies, A., Ferguson, M., Katz, K. & Schasberger, B. (1994). The Penn treebank: Annotating predicate argument structure. *Proceedings of the ARPA Human Language Technology Workshop*.

Mathews, A., Fox, E., Yiend, J. & Calder, A. (2003). The face of fear: Effects of eye gaze and emotion on visual attention. *Visual Cognition* 10, 823-835.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University of Chicago Press.

McNeill, D. (2005). *Gesture & Thought.* Chicago: University of Chicago Press.

McNeill, D. (2006). Gesture, Gaze, and Ground. *Lecture Notes in Computer Science* 3869, 1-14.

McNeill, D. (2008). Unexpected metaphors. In: A. Cienki & C. Müller (Eds.), *Metaphor and Gesture*, 155-170, Amsterdam/Philadelphia: Benjamins.

Michaelis, L. (2006). Construction Grammar. In: K. Brown (Ed.), *The Encyclopedia of Language and Linguistics*, Second edition, volume 3, 73-84, Oxford: Elsevier.

Michelas, A. & Nguyen, N. (2012). Speech imitation between speakers influences the realization of initial rises in French intonation. *Proceedings of the International Symposium on Imitation and Convergence in Speech*, 973-976.

Mol, L., Krahmer, E., Maes, A. & Swerts, M. (2012). Adaptation in gesture: Converging hands or converging minds? *Journal of Memory and Language* 66, 249-264.

Mondada, L. (2007). Multimodal resources for turn-taking : pointing and the emergence of possible next speakers. *Discourse Studies* 9, 194-225.

Montgomery, K., Isenberg, N. & Haxby, J. (2007). Communicative hand gestures and object-directed hand movements activated the mirror neuron system. *Social Cognitive and Affective Neuroscience* 2, 114-122.

Muir, L. & Richardson, I. (2005). Perception of sign language and its application to visual communications for deaf people. Journal of Deaf Studies and Deaf Education 10, 390-40.

Nass, C. & Lee, K. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology* 7, 171–181.

Neider, M., Chen, X., Dickinson, C., Brennan, S. & Zelinsky, G. (2010). Coordinating spatial referencing using shared gaze. *Psychonomic Bulletin & Review* 17, 718-724.

Nenkova, A., Gravano, A. & Hirschberg, J. (2008). High frequency word entrainment in spoken dialogue. *Proceedings of Association for Computational Linguistic 08*, 169-172.

Newman-Norlund, R., Van Schie, H., Van Zuijlen, A. & Bekkering, H. (2007). The mirror neuron system is more active during complementary compared with imitative action. *Nature Neuroscience* 10, 817-818.

Nielsen, G. (1962). *Studies in self confrontation*. Copenhagen: Munksgaard.

Novick, D., Hansen, B. & Ward, K. (1996) Coordinating turn-taking with gaze. *Proceedings of the international conference on spoken language processing*, 1888-1891.

Oben, B. & Brône, G. (forthc.). What you see is what you do. On the relationship between gaze and gesture in multimodal alignment. *Language and Cognition*.

Oertel C., Wlodarczak M., Edlund J., Wagner P. & Gustafson J. (2012). Gaze patterns in turn-taking. *Proceedings of Interspeech*, 2247-2250.

Oostdijk, N. (2000). The Spoken Dutch Corpus. Overview and first Evaluation. *Proceedings of the 2nd International Conference on Language Resources and Evaluation*.

Östman J.-O. & Fried M. (2005). The cognitive grounding of Construction Grammar. In: J.-O.Östman & M. Fried (Eds.), *Construction Grammars? Cognitive grounding and theoretical extensions*, 1-16, Amsterdam: John Benjamins.

Paggio, P., Allwood, J., Ahlsén, E., Jokinen, K. & Navarretta, C. (2010). The NOMCO Multimodal Nordic Resource - Goals and Characteristics. *Proceedings of the 7th International Conference on Language Resources and Evaluation*, 2968-2973.

Paggio, P., Heylen, D. & Kipp, M. (Eds.) (2013). Multimodal Corpora [Special issue]. *Journal on Multimodal User Interfaces* 7, 1-170.

Pardo, J. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America* 119, 2382-2393.

Parrill, F. & Kimbara, I. (2006). Seeing and hearing double: The influence of mimicry in speech and gesture on observers. *Journal of Nonverbal Behavior* 30, 157-166.

Pearson, J., Hu, J., Branigan, H., Pickering, M. & Nass, C. (2006). Adaptive language behaviour in HCI: How expectations and beliefs about a system affect users' word choice. *Proceedings of the 2006 Conference on Human Factors in Computing Systems*, 1177-1180.

Pennebaker, R., Booth, R. & Francis, M. (2007). Linguistic Inquiry and Word Count: LWIC.

Perrin, L., Deshaies, D. & Paradis, C. (2003). Pragmatic functions of local diaphonic repetitions in conversation. *Journal of Pragmatics* 35, 1843-1860.

Piaget, J. (1932). *The moral judgment of the child.* Glencoe: The Free Press.

Piazza, J., Bering, J. & Ingram, G. (2011). Princess Alice is watching you: Children's belief in an invisible person inhibits cheating. *Journal of Experimental Child Psychology* 109, 311-320.

Pickering, M. & Ferreira, V. (2008). Structural priming: A critical review. *Psychological Bulletin* 134, 427-459.

Pickering, M. & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences* 36, 329-347.

Pickering, M. & Garrod, S. (2004). Towards a Mechanistic Psychology of Dialogue. *Behavioural and Brain Sciences* 27, 169-225.

Pickering, M. & Garrod, S. (2006). Alignment as the Basis for Successful Communication. *Research on Language and Communication* 4, 203-288.

Pine, K., Lufkin, N. & Messer, D. (2004). More gestures than answers: Children learning about balance. *Developmental Psychology* 40, 1059-1067.

Porzel, R., Scheffler, A., & Malaka, R. (2006). How entrainment increases dialogical effectiveness. *Proceedings of the IUI'06 Workshop on Effective Multimodal Dialogue Interaction*, 1-8.

Posner, M., Snyder, C. & Davidson, B. (1980). Attention and the detection of signals. *Journal of Experimental Psychology* 109, 160–174.

Postma, M., Brunninkhuis, N. & Postma, E. (2013). Eye Gaze Affects Vocal Intonation Mimicry. *Proceedings of the 35th Annual Conference of the Cognitive Science Society,* 1139-1144.

Powell, L., Roberts, G. & Nettle, D. (2012). Eye images increase charitable donations: Evidence from an opportunistic field experiment in a supermarket. *Ethology* 118, 1-6.

Reitter, D., Moore, D. & Keller, F. (2006). Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. Proceedings of the 28th Annual Conference of the Cognitive Science Society, 685-690.

Richardson, D. & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. Cognitive Science 29, 1045-1060.

Richardson, D., Dale, R., & Kirkham, N. (2007). The art of conversation is coordination. Common ground and the coupling of eye movements during dialogue. *Psychological Science* 18, 407-413.

Richardson, D., Dale, R. & Tomlinson, J. (2009). Conversation, gaze coordination, and beliefs about visual context. *Cognitive Science* 33, 1468-1482.

Riordan, M., Dale, R., Kreuz, R. & Olney, A. (2011). Evidence for alignment in a computer-mediated text-only environment. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 2411-2416.

Rizzolatti, G., Fogassi, L. & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews: Neuroscience* 2, 661-670.

Roche, J., Dale, R. & Caucci, G. (2010). Doubling up on double meanings: Pragmatic alignment. *Language and Cognitive Processes* 27, 1-24.

Sakita, T. (2006). Parallelism in conversation. Resonance, schematization, and extension from the perspective of dialogic syntax and cognitive linguistics. *Pragmatics & Cognition* 14, 467-500.

Selting, M., Auer, P. et al. (2009). Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion* 10, 353-402.

Shaik, J., van Baaren, R., Bekkering, H. & Hunnius, S. (2013). Evidence for nonconscious behavior-copying in young children. *Proceedings of the 35th Annual Conference of the Cognitive Science Society,* 1516-1521.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal* 27, 379-423.

Shockley, K., Santana, M. & Fowler, C. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance* 29, 326–332.

Stanford, G. & Webster, S. (1996). A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. *Journal of Personality and Social Psychology* 70, 1231-1240.

Stel, M., Van Baaren, R., Blascovich, J., van Dijk, E., McCall, C., Pollmann, M., van Leeuwen, M., Mastop, J. & Vonk, R. (2010). Effects of a priori liking on the elicitation of mimicry. *Experimental Psychology* 57, 412-418.

Streeck, J. (2009). *Gesturecraft - The manufacture of meaning*. Amsterdam/ Philadelphia: John Benjamins.

Streeck, J. (2008). Depicting by gesture. *Gesture* 8, 285-301.

Szczepek Reed, B. (2010). Prosody and alignment: A sequential perspective. *Cultural Studies of Science Education* 5, 859-867.

Szmrecsanyi, B. (2005). Language users as creatures of habit: a corpus-linguistic analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1, 113-150

Tannen, D. (1987). Repetition in conversation: Toward a poetics of talk. *Language* 63, 574-605.

Tannen, D. (1989). *Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse.* Cambridge: Cambridge University Press.

Tollefsen, D. & Dale, R. (2012). Naturalizing joint action: A process-based approach. *Philosophical Psychology* 25, 385-407.

Tolston, M., Ariyabuddhiphongs, K., Riley, M. & Shockley, K. (2014). Cross-recurrence quantification analysis of the influence of coupling constraints on interpersonal coordination and communication. *Springer Proceedings in Mathematics & Statistics* 103, 157-171.

Tomasello, M., Hare, B., Lehmann, H. & Call, J. (2007). Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. *Journal of Human Evolution* 52, 314-320.

Uldall, B., Hall, C., & Chartrand, T. (in prep.) Optimal distinctiveness theory and mimicry: When being too distinct leads to more mimicry of others.

Van Baaren, R., Janssen, L., Chartrand, T. & Dijksterhuis, A. (2009). Where is the love? Social aspects of mimicry. *Philosophical Transactions of the Royal Society B, Biological sciences* 364, 2381–2389.

Van Baaren, R., Holland, R., Kawakami, K. & van Knippenberg, A. (2004). Mimicry and pro-social behavior. *Psycholigal Science* 15, 71–74.

Van den Bosch, A., Busser, G., Daelemans, W. & Canisius, S. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, 99-114.

Van Engen, K., Baese-Berk, M., Baker, R., Choi, A., Kim, M. & Bradlow, A. (2010). The Wildcat corpus of native- and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and speech* 53, 510-540.

Van Son, R., Wesseling, W., Sanders, E. & Van Der Heuvel, H. (2008). The IFADV corpus: A free dialog video corpus. *Proceedings of the 6th International Conference on Language Resources and Evaluation*, 501-*508.*

Vaughan, B. (2011). Prosodic synchrony in co-operative task-based dialogues: a measure of agreement and disagreement. *Proceedings of Interspeech,* 1865-1868.

Venables, W. & Ripley, B. (2002). *Modern Applied Statistics with S. Fourth Edition*. Berlin: Springer.

Vertegaal, R., Slagter, R., Van der Veer, G. & Nijholt, A. (2001). Eye Gaze Patterns in Conversations: There is More to Conversational Agents Than Meets the Eyes. *Proceedings of the Conference on Human Factors in Computing Systems*, 301-308.

Vogels, J. (2014). Referential choices in language production (Unpublished doctoral dissertation). Tilburg University, Tilburg.

Vorwerg, C. (2013). Language variation and mutual adaptation in interactive communication: Putting together psycholinguistic and sociolinguistic perspectives. In: I. Wachsmuth, J. de Ruiter, S. Kopp, & P. Jaecks (Eds.), *Advances in Interaction Studies. Alignment in Communication: Towards a New Theory of Communication*. Amsterdam: Benjamins, 149-166.

Wang, Y., Newport, R. & Hamilton, A. (2011). Eye contact enhances mimicry of intransitive hand movements. *Biology Letters* 7, 7–10.

Wang, Y. & Hamilton, A. (2014). Why does gaze enhance mimicry? Placing gaze-mimicry effects in relation to other gaze phenomena. *The Quarterly Journal of Experimental Psychology* 67, 747-762.

Webb, J.T., 1972. Interview synchrony: an investigation of two speech rate measures in an automated standardized interview. In: W. Siegman & B. Pope (Eds.), *Studies in Dyadic Communication*. New York: Pergamon Press, 115-133.

Weiner, E. & Labov, W. (1983). Constraints on the agentless passive. *Journal* of Linguistics 19, 29-58.

Zima, E. (2013). *Kognition in der Interaktion. Eine kognitiv-linguistische und gesprächsanalytische Studie dialogischer Resonanz in österreichischen Parlamentsdebatten*. Heidelberg: Universitätsverlag Winter.

Samenvatting

Geeuwen werkt aanstekelijk. Als je iemand ziet geeuwen, verhoogt de kans sterk dat je zelf ook zal geeuwen. Dit soort van kopiegedrag stellen we ook op talig vlak vast. Mensen kopiëren schaamteloos de woorden, syntactische constructies, intonatie, uitspraak, etc. van hun gesprekspartners. Soms gebeurt dat bewust (bijvoorbeeld om een humoristisch effect te genereren) en soms onbewust (bijvoorbeeld wanneer blijkt dat je dezelfde lichaamshouding dan je gesprekspartner hebt aangenomen). Het feit *dat* mensen elkaars taal en gedrag kopiëren wanneer ze met elkaar interageren, staat vast. Over de reden *waarom* ze dat doen, is er nog veel discussie. Met dit doctoraat willen we een indirecte bijdrage leveren aan die waarom-vraag. De bijdrage is indirect omdat we een empirische studie doen naar *hoe* en *wanneer* mensen elkaar kopiëren. Een beter inzicht in die hoe- en wanneer-aspecten kan ons ultiem meer leren over waarom sprekers elkaar massaal nabootsen.

Met ons onderzoek benaderen we het fenomeen van kopiegedrag langs twee invalshoeken: een multimodaal en temporeel perspectief. Met dat eerste willen we kopiegedrag op verschillende gedragskanalen (i.c. taal, gebaren en oogbewegingen) in beeld brengen. Op welke manier is kopiegedrag op het ene niveau gelinkt aan dat op een ander niveau? Bijvoorbeeld, als mensen elkaars woorden kopiëren, wordt het bijhorende handgebaar dan ook gekopieerd? En welke rol speelt kijkgedrag in het verklaren van kopiegedrag: kopiëren mensen elkaars woorden of gebaren vaker als ze worden aangekeken door hun gesprekspartner? Of als ze zelf expliciet op het gebaar van die gesprekspartner gefocust hebben?

De tweede invalshoek in ons onderzoek is een temporeel perspectief. Gesprekspartners kopiëren elkaar niet constant, dus willen we weten wanneer de frequentie van het kopiegedrag het hoogst is. Zien we een toename van die frequentie doorheen de tijd? Klopt het met andere woorden dat sprekers elkaar meer kopiëren naarmate ze langer met elkaar

spreken? Of zien we heel lokale pieken van kopiegedrag, afgewisseld met de afwezigheid ervan? Of kunnen we nog een ander temporeel patroon vaststellen?

Om een antwoord te vinden op bovenstaande vragen hebben we vijf case studies uitgewerkt. In een eerste case studie bestudeerden we de rol van kijkgedrag op kopiegedrag van woorden en van handgebaren. We vonden dat sprekers elkaars woorden significant vaker kopiëren wanneer hun gesprekspartner hen aankijkt, dan wanneer die gesprekspartner hen niet aankijkt. Meer preciezer geformuleerd: als spreker1 een woord gebruikt terwijl hij spreker2 aankijkt, is de kans groter dat spreker2 dat woord ook zal gebruiken, vergeleken met wanneer spreker1 een woord gebruikt zonder spreker2 aan te kijken. Bij handgebaren bleek dat aankijkeffect er niet te zijn. Wat we daar wel vonden, was dat als spreker2 focust op het gebaar van spreker1, de kans significant groter is dat hij dat gebaar ook zal kopiëren, vergeleken met wanneer spreker2 niet focust op het gebaar van spreker1.

In een tweede case studie hebben we onderzocht of dezelfde factoren kopiegedrag van woorden en van gebaren verklaren. We keken onder andere of het tijdsverschil tussen het gedrag van spreker1 en spreker2, het aantal keer dat de sprekers een bepaald woord of gebaar al gebruikt hadden, de hoeveelheid verstreken gesprekstijd, enz. een invloed hebben op het kopiegedrag van woorden en gebaren. In onze data bleek dat die twee niveaus door verschillende factoren worden verklaard. Voor gebaren bleek vooral de afstand tussen het gebaar van spreker1 en spreker2 belangrijk. Als sprekers twee gebaren tegelijkertijd gebruiken, blijken dat vaak dezelfde gebaren te zijn. Daarnaast zagen we ook een effect van gesprekstijd: hoe langer sprekers met elkaar interageren, hoe meer kopiegedrag van gebaren we vaststellen. Voor kopiegedrag van woorden was er maar één relevante factor: hoe vaker spreker1 een bepaald woord gebruikt, hoe hoger de kans dat spreker2 dat woord ook zal gebruiken.

In de derde case studie stond kijkgedrag centraal. De hoofdvraag was daar of sprekers hun kijkgedrag synchroniseren met het kijkgedrag of met het spreekgedrag van hun partner. Met andere woorden, kijkt spreker1

naar spreker2 omdat spreker2 aan het spreken is, of omdat spreker2 zelf naar spreker1 aan het kijken is? Onze resultaten toonden aan dat beiden het geval zijn. Sprekers synchroniseren hun kijkgedrag zowel met het kijkgedrag van hun partner als met het spreekgedrag van hun partner. Ze doen dat heel precies: van zodra een spreker naar het gezicht van de andere kijkt, beantwoord die andere dat door naar de ene te kijken. In de interactie tussen kijken en spreken vonden we een kleine vertraging: gemiddeld 0.3 seconden nadat iemand aan het spreken is, kijkt de luisteraar naar zijn sprekende gesprekspartner. Dit effect van synchronisatie van kijkgedrag bleek ook toe te nemen over gesprekstijd: hoe langer mensen met elkaar converseerden, hoe sterker de synchronisatie.

In case studie 4 bestudeerden we hoe de hoeveelheid kopiegedrag varieert doorheen langere gesprekken. We voerden eenzelfde studie uit voor een hele reeks van gedragsniveaus: toonhoogte, luidheid, spreeksnelheid, woorden, syntactische constructies en gebaren. Voor sommige niveaus vonden we een toename van kopiegedrag over gesprekstijd (toonhoogte, luidheid, functiewoorden, syntactische constructies en gebaren); voor sommige niveaus niet (spreeksnelheid, inhoudswoorden). Belangrijker is onze vaststelling dat die toename hoegenaamd niet geleidelijk is. Kopiegedrag blijkt heel dynamisch te variëren doorheen een gesprek. Soms kopiëren sprekers elkaars gedrag integraal en soms doen ze dat helemaal niet.

In de vijfde en laatste case studie was de vraag of kopiegedrag op het ene niveau samenvalt met kopiegedrag op andere niveaus. Enkel binnen het talige niveau vonden we significante correlaties. De momenten waarop sprekers elkaar qua toonhoogte kopiëren, doen ze dat ook voor luidheid. Dezelfde relatie vonden we voor functiewoorden en grammaticale constructies. Echter, tussen het talige niveau, gebaren en kijkgedrag konden we geen correlaties vaststellen. Blijkbaar loopt kopiegedrag op die niveaus behoorlijk onafhankelijk van elkaar.

Met dit doctoraat hebben we aangetoond dat kopiegedrag een erg dynamisch fenomeen is. Sprekers kopiëren elkaar niet constant, en de hoeveelheid kopiegedrag fluctueert erg sterk doorheen een gesprek. We zagen ook dat kopiegedrag op het ene niveau (talig) vrij los staat van

kopiegedrag op een ander niveau (gebaren): verschillende factoren verklaren het kopiegedrag op beide niveaus, en kopiegedrag op beide niveaus komt niet systematisch op dezelfde momenten tijdens een gesprek voor. Theoretisch gezien hebben we met dit werk een bijdrage geleverd aan de discussie over hoe automatisch en mechanistisch kopiegedrag is. Onze data tonen aan dat dat niet uitsluitend het geval is: kopiegedrag is erg dynamisch en contextafhankelijk en kan niet zomaar door één factor verklaard worden. Op methodologisch vlak hebben we vooruitgang geboekt in het objectief meten van de hoeveelheid kopiegedrag. Met die methodes kunnen we in vervolgonderzoek aan de slag om te bestuderen met welke sociale, cognitieve of interactionele factoren het geobserveerde kopiegedrag overeenkomt. Die kennis kan dan op zich weer een beter antwoord bieden op de vraag *waarom* mensen zo structureel gedrag van elkaar overnemen als ze met elkaar spreken.